Christoph Schiller

Motion Mountain

The Adventure of Physics

www.motionmountain.net

Christoph Schiller

Motion Mountain

The Adventure of Physics

available at
www.motionmountain.net

Editio decima octava.

Eighteenth revision, September 2005.

About the cover photograph, see page 1188.

To T.

τῷ ἐμοὶ δαὶμονι

Die Menschen stärken, die Sachen klären.

Contents

Detailed Contents

Second Part : Quantum Theory
What Is Matter? What Are Interactions?

Preface

Primum movere, deinde docere.

The intensity with which small children explore their environment suggests that there is a drive to grasp the way the world works, a 'physics instinct', built into each of us. What would happen if this drive, instead of dying out with the end of school education, were allowed to thrive in an environment without bounds, reaching from the atoms to the stars? Probably each adolescent would know more about nature than most senior physics teachers today. This text tries to provide this possibility to the reader. It acts as a guide in such an exploration, free of all limitations, of the world of *motion*. The project is the result of a threefold aim I have pursued since 1990: to present the basics of motion in a way that is simple, up to date and vivid.

In order to be *simple*, the text focuses on concepts and their understanding, while reducing the mathematics to the necessary minimum. Learning the concepts of physics is given precedence over using formulae in calculations. All topics are within the reach of an undergraduate. For the main domains of physics, the simplest summaries possible are presented. It is shown that physics describes motion in three steps. First there is everyday physics, or classical continuum physics. In the second step each domain of physics is based on an inequality for the main observable. Indeed, statistical thermodynamics limits entropy by $S \geqslant k/2$, special relativity limits speeds by $v \leqslant c$, general relativity limits force by $F \leqslant c^4/4G$, quantum theory limits action by $L \geqslant \hbar/2$ and quantum electrodynamics limits change of charge by $\Delta q \geqslant e$. By basing these domains of physics on limit principles, a simple, rapid and intuitive introduction is achieved. It is shown that the equations of each domain follow from the corresponding limit. The third step of physics is the unification of all these limits in a single description of motion. This way to learn physics should reward the curiosity of every reader – whether student or professional.

In order to be *up-to-date*, the text includes quantum gravity, string theory and M theory. But also the standard topics – mechanics, electricity, light, quantum theory, particle physics and general relativity – are greatly enriched by many gems and research results that are found scattered throughout the scientific literature.

In order to be *vivid*, a text wants to challenge, to question and to dare. This text tries to startle the reader as much as possible. Reading a book on general physics should be similar to a visit to a magic show. We watch, we are astonished, we do not believe our eyes, we think and finally – maybe – we understand the trick. When we look at nature, we often have the same experience. The text tries to intensify this by following a simple rule: on each page, there is at least one surprise or one provocation to think about. Numerous challenges are proposed. All are as interesting as possible. Hints or answers are given in the appendix.

A surprise has the strongest effect whenever it questions everyday observations. In this text most surprises are taken from daily life, in particular, from the experiences one makes when climbing a mountain. Observations about trees, stones, the Moon, the sky and people are used wherever possible; complex laboratory experiments are mentioned only where necessary. All surprises are organized to lead in a natural way to the most extreme conclusion of all, namely that continuous space and time do not exist. These concepts, useful as they may be in everyday life, are only approximations that are not valid in the general case. Time and space turn out to be mental crutches that *hinder* the complete exploration of the world.

Enjoying curiosity to full intensity and achieving freedom of thought leads to a strong and dependable character. Indeed, exploring a limit requires courage. Courage is also needed to drop space and time as tools for the description of the world. Changing thinking habits produces fear; but nothing is more intense and satisfying than overcoming one's own fears. Achieving a description of the world without the use of space and time may be the most beautiful of all adventures of the mind.

Eindhoven and other places, 26 September 2005

CS

A request

In exchange for getting this text for free, please send a short email on the following issues:

- What was unclear?
- What should be improved? Challenge 1 ny
- What did you miss?

Material on the specific points listed on the http://www.motionmountain.net/project.html web page is most welcome of all. Thank you in advance for your input, also in the name of all other readers. For a particularly useful contribution you will be mentioned in the acknowledgements, receive a reward, or both. But above all, enjoy the reading.

C. Schiller
fb@motionmountain.net

Acknowledgements

Many people who have kept their gift of curiosity alive have helped to make this project come true. Most of all, Saverio Pascazio has been – present or not – a constant reference for this project. Also Fernand Mayné, Anna Koolen, Ata Masafumi, Roberto Crespi, Luca Bombelli, Herman Elswijk, Marcel Krijn, Marc de Jong, Martin van der Mark, Kim Jalink, my parents Peter and Isabella Schiller, Mike van Wijk, Renate Georgi, Paul Tegelaar, Barbara and Edgar Augel, M. Jamil, Ron Murdock, Carol Pritchard and, most of all, my wife Britta have provided valuable advice and encouragement.

The project and the collection of material owes to many. Most useful was the help of Mikael Johansson, Bruno Barberi Gnecco, Lothar Beyer, the numerous improvements by Bert Sierra, the detailed suggestions by Claudio Farinati, the many improvements by Eric Sheldon, the continuous help and advice of Jonatan Kelu, and in particular the extensive, passionate and conscientious help of Adrian Kubala.

Important material was provided by Bert Peeters, Anna Wierzbicka, William Beaty, Jim Carr, John Merrit, John Baez, Frank DiFilippo, Jonathan Scott, Jon Thaler, Luca Bombelli, Douglas Singleton, George McQuarry, Tilman Hausherr, Brian Oberquell, Peer Zalm, Martin van der Mark, Vladimir Surdin, Julia Simon, Antonio Fermani, Don Page, Stephen Haley, Peter Mayr, Allan Hayes, Norbert Dragon, Igor Ivanov, Doug Renselle, Wim de Muynck, Steve Carlip, Tom Bruce, Ryan Budney, Gary Ruben, Chris Hillman, Olivier Glassey, Jochen Greiner, squark, Martin Hardcastle, Mark Biggar, Pavel Kuzin, Douglas Brebner, Luciano Lombardi, Franco Bagnoli, Lukas Fabian Moser, Dejan Corovic, Paul Vannoni, John Haber, Saverio Pascazio, Klaus Finkenzeller, Leo Volin, Jeff Aronson, Roggie Boone, Lawrence Tuppen, Quentin David Jones, Arnaldo Uguzzoni, Frans van Nieuwpoort, Alan Mahoney, Britta Schiller, Petr Danecek, Ingo Thies, Vitaliy Solomatin, Carl Offner, Nuno Proença, Elena Colazingari, Paula Henderson, Daniel Darre, Wolfgang Rankl, John Heumann, Joseph Kiss, Martha Weiss, Antonio González, Antonio Martos, John Heumann, André Slabber, Ferdinand Bautista, Zoltán Gácsi, Pat Furrie, Michael Reppisch, Enrico Pasi, Thomas Köppe, Martin Rivas, Herman Beeksma, Tom Helmond, John Brandes, Vlad Tarko, Nadia Murillo, Ciprian Dobra, Romano Perini, Harald van Lintel, Andrea Conti, François Belfort, Dirk Van de Moortel, Heinrich Neumaier, Jarosław Królikowski, John Dahlmann and all those who wanted to remain unnamed.

The software tools were refined with the extended help on fonts and typesetting by Michael Zedler and Achim Blumensath and with the repeated and valuable support of Donald Arseneau; help came also from Ulrike Fischer, Piet van Oostrum, Gerben Wierda, Klaus Böhncke, Craig Upright, Herbert Voss, Andrew Trevorrow, Danie Els, Heiko Oberdiek, Sebastian Rahtz, Don Story, Vincent Darley, Johan Linde, Joseph Hertzlinger, Rick Zaccone and John Warkentin.

Page 1188

Many illustrations that shape this text were made available by the copyright holders. A warm thank you to all of them; they are mentioned in the dedicated credit section in Appendix E. In particular, Luca Gastaldi and Antonio Martos produced specific images for this text. Both the book and the website owe most to the suggestions and support of my wife Britta.

1. An appetizer

Die Lösung des Rätsels des Lebens in Raum und Zeit
liegt *außerhalb* von Raum und Zeit.*
Ludwig Wittgenstein, *Tractatus*, 6.4312

What is the most daring, amazing and exciting journey we can make in a lifetime? What is the most interesting place to visit? We can travel to places that are as remote as possible, like explorers or cosmonauts, we can look into places as far away as we can imagine, like astronomers, we can visit the past, like historians or archaeologists, or we can delve as deeply as possible into the human soul, like artists or psychologists. All these voyages lead either to other places or to other times (or nowadays, to other servers on the internet). However, we can do better.

The most daring trip is not the one leading to the most inaccessible place, but the trip leading to where there is no place at all. Such a journey implies leaving the prison of space and time and venturing beyond it, into a domain where there is no position, no present, no future and no past, where we are free of all restrictions, but also of any security of thought. There, discoveries are still to be made and adventures to be fought. Almost nobody has ever been there; humanity has so far taken 2500 years for the trip and still has not completely achieved it.

To venture into this part of nature, we need to be curious about the essence of travel itself, and in particular about its details and its limits. The essence of any travel is *motion*. By exploring motion we will be lead to the most fascinating adventures in the universe.

The quest to understand motion in all its details and limitations can be pursued behind a desk, with a book, some paper and a pen. But to make the adventure more apparent, this text tells the story of the quest as the ascent of a mountain. Every step towards the top corresponds to a step towards higher precision in the description of motion. In addition, each step will increase the pleasure and the encountered delights. At the top of the mountain we shall arrive in the domain we were looking for, where 'space' and 'time' are words that have lost all meaning and where the sight of the world's beauty is overwhelming and unforgettable.

Thinking without time or space is difficult but fascinating. In order to get a taste of the issues involved, try to answer the following questions without ever referring to either space or time:**

Challenge 2 n

- Can you *prove* that two points extremely close to each other always leave room for a third point in between?
- Can you describe the shape of a knot over the telephone?
- Can you explain on the telephone what 'right' and 'left' mean, or what a mirror is?
- Have you ever tried to make a telephone appointment with a friend without using any time or position term, such as clock, hour, place, where, when, at, near, before, after, near, upon, under, above, below?
- Can you describe the fall of a stone without using space or time?

* The solution of the riddle of life in space and time lies *outside* space and time.

** Solution to *challenges* are either given on page 1134 or later on in the text. Challenges are classified as research level (r), difficult (d), normal student level (n) and easy (e). Challenges with no solution yet are marked (ny).

- Do you know of *any observation at all* that you can describe without concepts from the domains 'space', 'time' or 'object'?
- Can you explain what time is? And what clocks are?
- Can you imagine a finite history of the universe, but without a 'first instant of time'?
- Can you imagine a domain of nature where matter and vacuum are indistinguishable?
- Have you ever tried to understand why motion exists?

This book tells how to achieve these and other feats, bringing to completion an ancient dream of the human spirit, namely the quest to describe *every* possible aspect of motion.

Why do your shoestrings remain tied? They do so because space has three dimensions. Why not another number? Finding the answer has required the combined effort of researchers over thousands of years. The answer was only found by studying motion up to its smallest details and by exploring each of its limits.

Why do the colours of objects differ? Why does the Sun shine? Why does the Moon not fall out of the sky? Why is the sky dark at night? Why is water liquid but fire is not? Why is the universe so big? Why can birds fly but men can't? Why is lightning not straight? Why are atoms neither square, nor the size of cherries? These questions seem to have little in common; but that impression is wrong. They are all about motion – about its details and its limitations. Indeed, they all appear and are answered in what follows. Studying the limits of motion we discover that when a mirror changes its speed it emits light. We also discover that gravity can be measured with a thermometer. We find that there are more cells in the brain than stars in the galaxy; people almost literally have a whole universe in their head. Exploring any detail of motion is already an adventure in itself.

By exploring the properties of motion we will find that in contrast to personal experience, motion never stops. We will find out why the floor cannot fall. We will understand why the speed of computers cannot be made arbitrary high. We will see that perfect memory cannot exist. We will understand that nothing can be perfectly black. We will learn that every clock has a certain probability of going backwards. We will discover that time literally does not exist. We will find out that all objects in the world are connected. We will learn that matter cannot be distinguished precisely from empty space. We will learn that we are literally made of nothing. We will learn quite a few things about our destiny. And we will understand why the world is not different from what it is.

Understanding motion, together with all its details and all its limits, implies asking and answering three specific questions.

How do things move? The usual answer states that motion is an object changing position over time. This seemingly boring statement encompasses general relativity, one of the most amazing descriptions of nature ever imagined. We find that space is warped, that light does not usually travel in a straight line and that time is not the same for everybody. We discover that there is a maximum force of gravity and that, nevertheless, gravity is not an interaction, but rather the change of time with position. We understand that the blackness of the sky at night proves that the universe has a finite age. We also discover that there is a smallest entropy in nature, which prevents us from knowing everything about a physical system. In addition, we discover the smallest electrical charge. These and other strange properties of motion are summarized in the first part of this text, whose topic is classical physics. It directly leads to the next question.

What are things? Things are composites of a few types of particles. In addition, all interactions and forces – those of the muscles, those that make the Sun burn, those that make the Earth turn, those that determine the differences between attraction, repulsion, indifference, friction, creation and annihilation – are made of particles as well. The growth of trees, the colours of the sky, the burning of fire, the warmth of a human body, the waves of the sea and the mood changes of people are all variations of motion of particles. This story is told in more detail in the second part of the text, that on quantum mechanics. Here we will learn that there is a smallest change in nature. This minimum value forces everything to keep constantly changing. In particular, we will learn that it is impossible to completely fill a glass of wine, that eternal life is impossible, and that light can be transformed into matter. If this is still boring, read about the substantial dangers you incur when buying a can of beans.

Page 886

The first two parts of this text can be summarized with the help of a few limit principles:

$$\begin{aligned}
&\text{statistical thermodynamics limits entropy:} && S \geqslant k/2 \\
&\text{special relativity limits speed:} && v \leqslant c \\
&\text{general relativity limits force:} && F \leqslant c^4/4G \\
&\text{quantum theory limits action:} && L \geqslant \hbar/2 \\
&\text{quantum electrodynamics limits charge:} && \Delta q \geqslant e\,.
\end{aligned} \tag{1}$$

We will see that each of the constants of nature $k/2$, c, $c^4/4G$, $\hbar/2$ and e that appears on the right side is also a *limit* value. We will discover that the equations of the corresponding domain of physics follow from this limit property. After these results, the path is prepared for the final theme of the mountain climb.

What are particles, position and time? The recent results of an age-long search are making it possible to start answering this question. One just needs to find a description which explains all limit principles at the same time. This third part is not complete yet, because the final research results are not yet available. Nevertheless, the intermediate results are challenging:

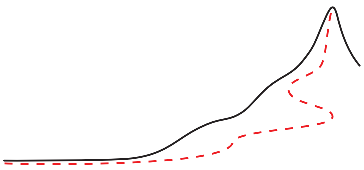
- It is known already that space and time are not continuous, that – to be precise – neither points nor particles exist, and that there is no way to distinguish space from time, nor vacuum from matter, nor matter from radiation.

Page 920

- It is known already that nature is not simply made of particles and vacuum, in contrast to what is often said.
- It seems that position, time and every particle are aspects of a complex, *extended* entity that is incessantly varying in shape.
- Mysteries that should be cleared up in the coming years are the origin of the three dimensions of space, the origin of time and the details of the big bang.
- Research is presently discovering that motion is an intrinsic property of matter and radiation and that, as soon as we introduce these two concepts in the description of nature, motion appears automatically. On the other hand, it is impossible *not* to introduce these concepts, because they necessarily appear when we divide nature into parts, an act we cannot avoid because of the mechanisms of our senses and therefore of our thinking.
- Research is also presently uncovering that the final description of nature, with complete precision, does not use any form of infinity. We find, step by step, that all infinities appearing in the human description of nature, both the infinitely large as well as the infinitely

small, result from approximations. ‘Infinity’ turns out to be an exaggeration that does not apply to nature at all. Then however, we find that the precise description does not include any finite quantities either! These and many other astonishing results of modern physics form the third part of this text.

This third and final part of the text thus develops the present state of the search for a unified description of general relativity and quantum mechanics. The secrets of space, time, matter and forces have to be unravelled to achieve it. It is a fascinating story, assembled piece by piece by thousands of researchers. At the end of the ascent, at the top of the mountain, the idea of motion will have undergone a complete transformation. Without space and time, the world will look magical, incredibly simple and astonishingly fascinating at the same time: pure beauty.

First Part 

Classical Physics: How Do Things and Images Move?

Where the experience of hiking and other motion
leads us to introduce, for its description,
the concepts of velocity, time, length, mass and charge,
as well as action, field and manifold,
allowing us to discover limits to
speed, entropy, force and charge,
and thus to understand – among other things –
why we have legs instead of wheels,
how empty space can bend, wobble and move,
what love has to do with magnets and amber,
and why we can see the stars.

Chapter I

Galilean Motion

2. Why should we care about motion?

All motion is an illusion.

Zeno of Elea*

Wham! The lightning striking the tree nearby violently disrupts our quiet forest walk and causes our hearts to suddenly beat faster. But the fire that started in the tree quickly fades away. The gentle wind moving the leaves around us helps to restore the calmness of the place. Nearby, the water in a small river follows its complicated way down the valley, reflecting on its surface the ever-changing shapes of the clouds.

Motion is everywhere: friendly and threatening, horrible and beautiful. It is fundamental to our human existence. We need motion for growing, for learning, for thinking and for enjoying life. We use motion for walking through a forest, for listening to its noises and for talking about all this. Like all animals, we rely on motion to get food and to survive dangers. Plants by contrast cannot move (much); for their self-defence, they developed *poisons*. Examples of such plants are the stinging nettle, the tobacco plant, digitalis, belladonna and poppy; poisons include caffeine, nicotine, curare and many others. Poisons such as these are at the basis of most medicines. Therefore, most medicines exist essentially because plants have no legs. Like all living beings, we need motion to reproduce, to breathe and to digest; like all objects, motion keeps us warm.

Figure 1 An example of motion observed in nature

Motion is the most fundamental observation about nature at large. It turns out that *everything* which happens in the world is some type of motion. There are no exceptions. Motion is such a basic part of our observations that even the origin of the word is lost in the darkness of Indo-European linguistic history. The fascination of motion has always made it a favourite object of curiosity. By the fifth century BCE in ancient Greece, its study had been given a name: *physics*. Ref. 1

Motion is also important to the human condition. Who are we? Where do we come from? What will we do? What should we do? What will the future bring? Where do people come from? Where do they go to? What is death? Where does the world come from? Where does life lead to? All these questions are about motion. The study of motion provides answers which are both deep and surprising.

* Zeno of Elea (*c.* 450 BCE), one of the main exponents of the Eleatic school oh philosophy.

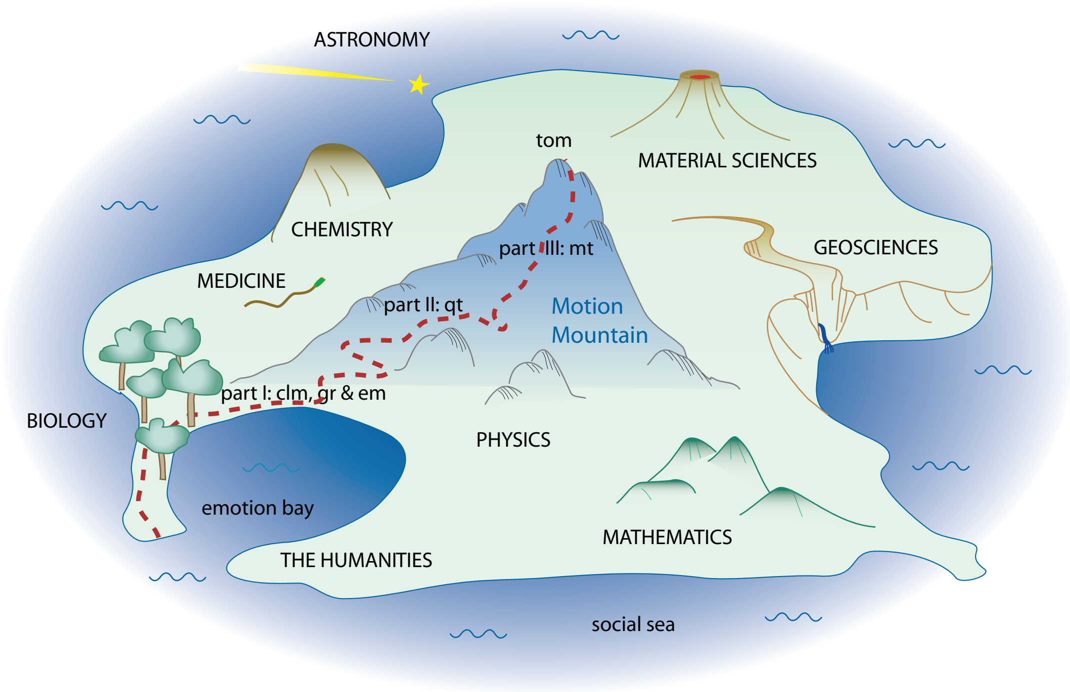


Figure 2 Experience Island, with Motion Mountain and the trail to be followed (clm: classical mechanics, gr: general relativity, em: electromagnetism, qt: quantum theory, mt: M-theory, tom: the theory of motion)

Motion is mysterious. Though found everywhere – in the stars, in the tides, in our eyelids – neither the ancient thinkers nor myriads of others in the following 25 centuries have been able to shed light on the central mystery: *what is motion?* We will discover that the standard reply, 'motion is the change of place in time', is inadequate. Just recently an answer has finally been found. This is the story of the way to reach it.

Motion is a part of human experience. If we imagine human experience as an island, then destiny, symbolized by the waves of the sea, carried us to its shore. Near the centre of the island an especially high mountain stands out. From its top we can oversee the whole landscape and get an impression of the relationships between all human experiences, in particular between the various examples of motion. This is a guide to the top of what I have called Motion Mountain. The hike is one of the most beautiful adventures of the human mind. Clearly, the first question to ask is:

Does motion exist?

> *Das Rätsel* gibt es nicht. Wenn sich eine Frage überhaupt stellen läßt, so *kann* sie beantwortet werden.*
> Ludwig Wittgenstein, *Tractatus*, 6.5

To sharpen the mind for the issue of motion's existence, have a look at Figure 3 and follow

* *The riddle* does not exist. If a question can be put at all, it *can* be answered.

Figure 3 Illusions of motion: look at the figure on the left and slightly move the page, or look at the white dot at the centre of the figure on the right and move your head back and forward

the instructions. In both cases the figures seem to rotate. How can one make sure that real motion is different from these or other similar illusions?* Challenge 3 n

Many scholars simply argued that motion does not exist at all. Their arguments deeply influenced the investigation of motion. Ref. 3 For example, the Greek philosopher Parmenides (born *c.* 515 BCE in Elea, a small town near Naples, in southern Italy) argued that since nothing comes from nothing, change cannot exist. He underscored the *permanence* of nature and thus consistently maintained that all change and thus all motion is an illusion. Ref. 4

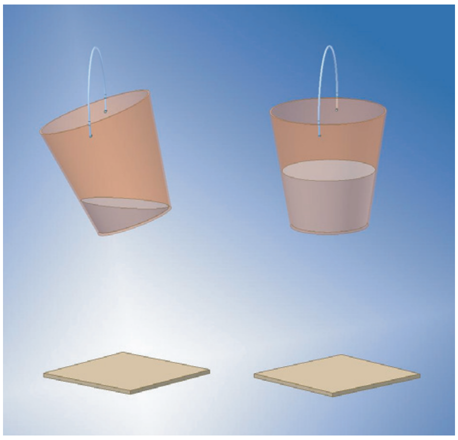
Heraclitus (*c.* 540 to *c.* 480 BCE) held the opposite view. He expressed it in his famous statement πάντα ῥεῖ 'panta rhei' or 'everything flows'.* He saw change as the essence of nature, in contrast to Parmenides. These two equally famous opinions induced many scholars to investigate in more detail whether in nature there are *conserved* quantities or whether *creation* is possible. We will uncover the answer later on; until then, you might ponder which option you prefer. Challenge 4 n

Parmenides' collaborator Zeno of Elea (born *c.* 500 BCE) argued so intensely against motion that some people still worry about it today. In one of his arguments he claims – in simple language – that it is impossible to slap somebody, since the hand first has to travel halfway to the face, then travel through half the distance that remains, then again so, and so on; the hand therefore should never reach the face. Zeno's argument focuses on the relation between *infinity* and its opposite, finitude, in the description of motion. In modern quantum theory, a similar issue troubles many scientists up to this day. Ref. 5

Zeno also maintained that by looking at a moving object at a *single* instant of time, one cannot maintain that it moves. Zeno argued that at a single instant of time, there is no difference between a moving and a resting body. He then deduced that if there is no difference at a single time, there cannot be a difference for longer times. Zeno therefore questioned whether motion can clearly be distinguished from its opposite, *rest*. Indeed, in the history of physics, thinkers switched back and forward between a positive and a negative answer. It was this very question that led Albert Einstein to the development of general relativity, one of the high points of our journey. We will follow the main answers

* Solutions to *challenges* are given either on page 1134 or later on in the text. Challenges are classified as research level (r), difficult (d), normal student level (n) and easy (e). Challenges with no solution yet are marked (ny).

* Appendix A explains how to read Greek text.



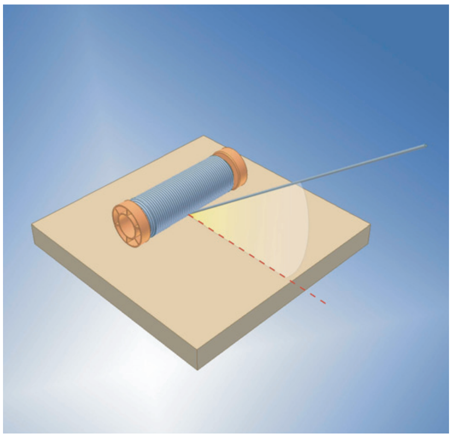


Figure 4 How much water is required to make a bucket hang vertically? At what angle does the pulled reel change direction of motion? (© Luca Gastaldi)

given in the past. Later on, we will be even more daring: we will ask whether single instants of time do exist at all. This far-reaching question is central to the last part of our adventure.

When we explore quantum theory, we will discover that motion is indeed – to a certain extent – an illusion, as Parmenides claimed. More precisely, we will show that motion is observed only due to the limitations of the human condition. We will find that we experience motion only because we evolved on Earth, with a finite size, made of a large but finite number of atoms, with a finite but moderate temperature, electrically neutral, large compared to a black hole of our same mass, large compared to our quantum mechanical wavelength, small compared to the universe, with a limited memory, forced by our brain to approximate space and time as continuous entities, and forced by our brain to describe nature as made of different parts. If any one of these conditions were not fulfilled, we would not observe motion; motion then would not exist. Each of these results can be uncovered most efficiently if we start with the following question:

How should we talk about motion?

Je hais le mouvement, qui déplace les lignes,
Et jamais je ne pleure et jamais je ne ris.
Charles Baudelaire, *La Beauté*.*

Like any science, the approach of physics is twofold: we advance with *precision* and with *curiosity*. Precision makes meaningful communication possible, and curiosity makes it worthwhile.** Whenever one talks about motion and aims for increased precision or for more detailed knowledge, one is engaged, whether knowingly or not, in the ascent of

* Charles Baudelaire (b. 1821 Paris, d. 1867 Paris) *Beauty*: 'I hate movement, which changes shapes, and never do I cry and never do I laugh.' The full text of this and the other poems from *Les fleurs du mal*, one of the finest books of poetry ever written, can be found at the http://hypermedia.univ-paris8.fr/bibliotheque/Baudelaire/Spleen.html website.

Ref. 6 ** For a collection of interesting examples of motion in everyday life, see the excellent book by Walker.

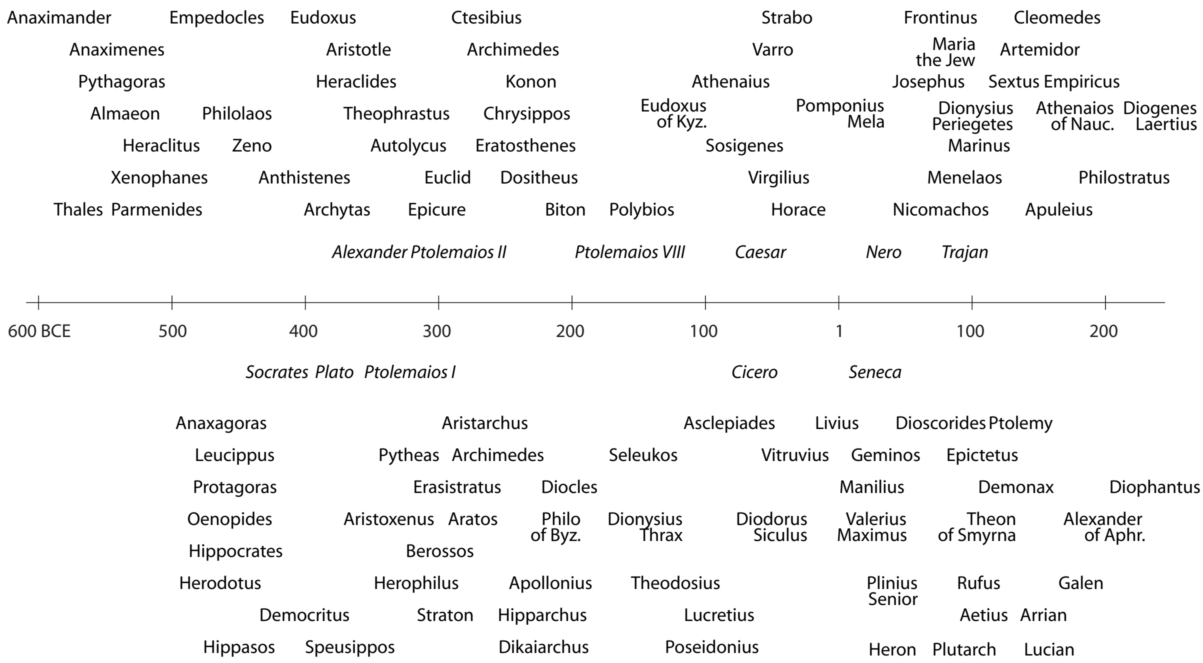


Figure 5 A time line of scientific and *political* personalities in antiquity (the last letter of the name is aligned with the year of death)

Motion Mountain. With every increase in the precision of description, one gains some height. The examples of Figure 4 make the point. When you fill a bucket with a little water, it does not hang vertically; if you continue adding water, it starts to hang vertically at a certain moment. How much water is necessary? When you pull a thread from a reel in the way shown, the reel will move either forwards or backwards, depending on the angle at which you pull. What is the limiting angle between the two possibilities?

Challenge 5 ny

High precision means going into fine details. This method actually *increases* the pleasure of the adventure.* The higher we get on Motion Mountain, the further we can see and the more our curiosity gets rewarded. The views offered are breathtaking, especially at the very top. The path we will follow – one of the many possible ones – starts from the side of biology and directly enters the forest lying at the foot of the mountain.

Ref. 7

Intense curiosity implies to go straight to the limits: understanding motion means to study the largest distances, the highest velocities, the smallest particles, the strongest forces and the strangest concepts. Let us start.

What are the types of motion?

> Every movement is born of a desire for change.

The best place to get a general overview on the types of motion is a big library; this is shown in Table 1. The domains in which motion, movements and moves play a role are

Challenge 6 n

* Distrust anybody who wants to talk you *out* of investigating details. He is trying to deceive you. Details are important. Be vigilant also during *this* walk.

Table 1 Content of books about motion found in a public library

Motion topics	Motion topics
motion pictures	motion therapy
motion perception Ref. 18	motion sickness
motion for fitness and wellness	motion for meditation
motion control in sport	motion ability as health check
perpetual motion	motion in dance, music and other arts
motion as proof of various gods Ref. 8	motion of stars and angels Ref. 9
economic efficiency of motion	emotion
motion as help to overcome trauma	motion in psychotherapy
locomotion of insects, horses and robots	commotion
motions in parliament	movements in art, sciences and politics
movements in watches	movements in the stock market
movement teaching and learning	movement development in children
musical movements	troop movements Ref. 10
religious movements	bowel movements
moves in chess	cheating moves in casinos Ref. 11
connection between gross national product and citizen mobility	

indeed varied. Already in ancient Greece people had the suspicion that all types of motion, as well as many other types of change, are related. It is usual to distinguish at least three categories.

The first category of change is that of material *transport*, such as a person walking or a leaf falling from a tree. Transport is the change of position and orientation of objects. To a large extent, the behaviour of people also falls into this category.

A second category of change groups observations such as the dissolution of salt in water, the formation of ice by freezing, the putrefaction of wood, the cooking of food, the coagulation of blood, and the melting and alloying of metals. These changes of colour, brightness, hardness, temperature and other material properties are all *transformations*. Transformations are changes not visibly connected with transport. To this category, a few ancient thinkers added the emission and absorption of light. In the twentieth century, these two effects were proven to be special cases of transformations, as were the newly discovered appearance and disappearance of matter, as observed in the Sun and in radioactivity. *Mind change change*, such as change of mood, of health, of education and of
Ref. 12 character, is also (mostly) a type of transformation.

Ref. 13 The third and especially important category of change is *growth*; it is observed for animals, plants, bacteria, crystals, mountains, stars and even galaxies. In the nineteenth century, changes in the population of systems, *biological evolution*, and in the twentieth century, changes in the size of the universe, *cosmic evolution*, were added to this category. Traditionally, these phenomena were studied by separate sciences. Independently they all arrived at the conclusion that growth is a combination of transport and transformation. The difference is one of complexity and of time scale.

Figure 6 An example of transport

At the beginning of modern science during the Renaissance, only the study of transport was seen as the topic of physics. Motion was equated to transport. The other two domains were neglected by physicists. Despite this restriction, the field of enquiry remains large, covering a large part of Experience Island. The obvious temptation is to structure the field by distinguishing types of transport by their origin. Movements such as those of the legs when walking are *volitional*, because they are controlled by one's will, whereas movements of external objects, such as the fall of a snowflake, which one cannot influence by will-power, are called *passive*. Children are able to make this distinction by about the age of six, and this marks a central step in the development of every human towards a precise description of the environment.* From this distinction stems the historical but now outdated definition of physics as the science of the motion of non-living things.

Then, one day, machines appeared. From that moment, the distinction between volitional and passive motion was put into question. Like living beings, machines are self-moving and thus mimic volitional motion. But careful observation shows that every part in a machine is moved by another, so that their motion is in fact passive. Are living beings also machines? Are human actions examples of passive motion as well? The accumulation of observations in the past 100 years made it clear that volitional movement* indeed has the same physical properties as passive motion in non-living systems. (Of

* Failure to pass this stage completely can result in various strange beliefs, such as in the ability to influence roulette balls, as found in compulsive players, or in the ability to move other bodies by thought, as found in numerous otherwise healthy-looking people. An entertaining and informative account of all the deception and self-deception involved in creating and maintaining these beliefs is given by JAMES RANDI, a professional magician, in *The Faith Healers*, Prometheus Books, 1989, as well as in several of his other books. See also his http://www.randi.org website for more details.

* The word 'movement' is rather modern; it was imported into English from the old French and became popular only at the end of the eighteenth century. It is never used by Shakespeare.

Figure 7 Transport, growth and transformation

course, from the emotional viewpoint, the differences are important; for example, *grace* can only be ascribed to volitional movements.) The distinction between the two types is thus not necessary and is dropped in the following. Since passive and volitional motion have the same properties, through the study of motion of non-living objects we can learn something about the human condition. This is most evident when touching the topics of determinism, causality, probability, infinity, time and sex, to name but a few of the themes we will encounter on the way.

Ref. 14

With the accumulation of observations in the nineteenth and twentieth centuries, even more restrictions on the study of motion were put into question. Extensive observations showed that all transformations and all growth phenomena, including behaviour change and evolution, are examples of transport as well. In other words, over 2 000 years of studies have shown that the ancient classification of observations was useless: all change is transport. In the middle of the twentieth century this culminated in the confirmation of an even more specific idea already formulated in ancient Greece: *every type of change is due to motion of particles.* It takes time and work to reach this conclusion, which appears only when one relentlessly pursues higher and higher precision in the description of nature. The first two parts of this adventure retrace the path to this result. (Do you agree with it?)

Challenge 7 n

The last decade of the twentieth century changed this view completely. The particle idea turns out to be wrong. This new result, already suggested by advanced quantum theory, is reached in the third part of our adventure through a combination of careful observation and deduction. But we still have some way to go before we reach there.

At present, at the beginning of our walk, we simply note that history has shown that classifying the various types of motion is not productive. Only by trying to achieve maximum precision can we hope to arrive at the fundamental properties of motion. Precision, not classification is the way to follow. As Ernest Rutherford said: 'All science is either physics or stamp collecting.'

To achieve precision in our description of motion, we need to select specific examples of motion and study them in full detail. It is intuitively obvious that the most precise description is achievable for the *simplest* possible examples. In everyday life, this is the case for the motion of any non-living, solid and rigid body in our environment, such as a stone thrown through the air. Indeed, like all humans, we learned to throw objects long
Ref. 15 before we learned to walk. Throwing is one of the first physical experiment we performed by ourselves.* During our early childhood, by throwing stones and similar objects until our parents feared for every piece of the household, we explored the perception and the properties of motion. We do the same.

Die Welt ist unabhängig von meinem Willen.*
Ludwig Wittgenstein, *Tractatus*, 6.373

Perception, permanence and change

Only wimps specialise in the general case; real scientists pursue examples.
Beresford Parlett

Human beings enjoy perceiving. Perception starts before birth, and we continue enjoying it as long as we can. That is why television, even when devoid of content, is so successful. During our walk through the forest at the foot of Motion Mountain we cannot avoid perceiving. Perception is first of all the ability to *distinguish*. We use the basic mental act of distinguishing in almost every instant of life; for example, during childhood we first learned to distinguish familiar from unfamiliar observations. This is possible in combination with another basic ability, namely the capacity to *memorize* experiences. Memory gives us the ability to experience, to talk and thus to explore nature. Perceiving, classifying and memorizing together form *learning*. Without any one of these three abilities, we could not study motion.

Children rapidly learn to distinguish *permanence* from *variability*. They learn to *recognize* human faces, even though faces never look exactly the same each time they are seen. From recognition of faces, children extend recognition to all other observations. Recognition works pretty well in everyday life; it is nice to recognize friends, even at night, and even after many beers (not a challenge). The act of recognition thus always uses a form of *generalization*. When we observe, we always have a general idea in our mind. We specify the main ones.

Every forest can remind us of the essence of perception. Sitting on the grass in a clearing of the forest at the foot of Motion Mountain, surrounded by the trees and the silence typical of such places, a feeling of calmness and tranquillity envelops us. Suddenly, something moves in the bushes; immediately our eyes turn and the attention focuses. The nerve cells that detect motion are part of the most ancient piece of our brain, shared with birds
Ref. 16 and reptiles: the brain stem. Then the cortex, or modern brain, takes over to analyse the

* The importance of throwing is also seen from the terms derived from it: in Latin, words like *subject* or 'thrown below', *object* or 'thrown in front', and *interjection* or 'thrown in between'; in Greek, it led to terms like *symbol* or 'thrown together', *problem* or 'thrown forward', *emblem* or 'thrown into', and – last but not least – *devil* or 'thrown through'.

* The world is independent of my will.

type of motion and to identify its origin. Watching the motion across our field of vision, we observe two invariant entities: the fixed landscape and the moving animal. After we recognize it as a deer, we relax again.

How did we distinguish between landscape and deer? Several steps in the eye and in the brain are involved. Motion plays an essential part in them, as is best deduced from the flip movie shown in the lower left corners of these pages. Each image shows only a rectangle filled with a mathematically-random pattern. But when the pages are scanned, one discerns a shape moving against a fixed background. At any given instant, the shape cannot be distinguished from the background; there is no visible object at any given instant of time. Nevertheless it is easy to perceive its motion.* Perception experiments such as this one have been performed in many variations. Among others it was found that detecting such a window is nothing special; flies have the same ability, as do, in fact, all animals which have eyes.

Ref. 17

The flip movie in the lower left corner, like many similar experiments, shows two central connections. First, motion is perceived only if an *object* can be distinguished from a *background* or *environment*. Many motion illusions focus on this point.** Second, motion is required to define both the object and the environment, and to distinguish them from each other. In fact, the concept of space is – among others – an abstraction of the idea of background. The background is extended; the moving entity is localized. Does this seem boring? It is not; just wait a second.

We call the set of localized aspects that remain invariant or permanent during motion, such as size, shape, colour etc., taken together, a (physical) *object* or a (physical) *body*. We will tighten the definition shortly, since otherwise images would be objects as well. In other words, right from the start we experience motion as a *relative* process; it is perceived in relation and in opposition to the environment. The concept of object is therefore also a relative concept. But the basic conceptual distinction between localized, isolable objects and the extended environment is not trivial or unimportant. First, it smells of a circular definition. (Do you agree?) This issue will keep us very busy later on. Second, we are so used to our ability of isolating local systems from the environment that we take it for granted. However, as we will see in the third part of our walk, this distinction turns out to be logically and experimentally impossible!*** Our walk will lead us to discover the reason for this impossibility and its important consequences. Finally, apart from moving entities and the permanent background, we need a third concept, as shown in Table 2.

Challenge 8 n

Page 939

Wisdom is one thing: to understand the thought which steers all things through all things.
Heraclitus of Ephesus

Ref. 19

* The human eye is rather good at detecting motion. For example, the eye can detect motion of a point of light even if the change of angle is smaller than what can be distinguished in fixed images. Details of this and similar topics for the other senses are the domain of perception research.

Ref. 18

** The topic of motion perception is full of interesting aspects. An excellent introduction is chapter 6 of the beautiful text by Donald D. Hoffman, *Visual Intelligence – How We Create What We See*, W.W. Norton & Co., 1998. His collection of basic motion illusions can be experienced and explored on the associated http://aris.ss.uci.edu/cogsci/personnel/hoffman/hoffman.html website.

*** Contrary to what is often read in popular literature, the distinction *is* possible in quantum theory. It becomes impossible only when quantum theory is unified with general relativity.

Table 2 Family tree of the basic physical concepts

motion
the basic type of change

parts	*relations*	*background*
permanent	variable	permanent
bounded	unbounded	extended
shaped	unshaped	measurable

objects	*images*	*states*	*interactions*	*phase space*	*space-time*
impenetrable	penetrable	global	local	composed	simple

The corresponding aspects:

mass	intensity	instant	source	dimension	curvature
size	colour	position	domain	distance	topology
charge	appearance	momentum	strength	volume	distance
spin	disappearance	energy	direction	subspaces	area
etc.	etc.	etc.	etc.	etc.	etc.

world – nature – universe – cosmos
the collection of all parts, relations and backgrounds

Does the world need states?

> Das Feste, das Bestehende und der Gegenstand sind Eins. Der Gegenstand ist das Feste, Bestehende; die Konfiguration ist das Wechselnde, Unbeständige.*
> Ludwig Wittgenstein, *Tractatus*, 2.027 - 2.0271

What distinguishes the various patterns in the lower left corners of this text? In everyday life we would say: the *situation* or *configuration* of the involved entities. The situation somehow describes all those aspects which can differ from case to case. It is customary to call the list of all *variable* aspects of a set of objects their *(physical) state of motion*, or simply their *state*.

The situations in the lower left corners differ first of all in *time*. Time is what makes opposites possible: a child is in a house and the same child is outside the house. Time describes and resolves this type of contradictions. But the state not only distinguishes situations in time. The state contains *all* those aspects of a system (i.e., of a group of objects) which set it apart from all *similar* systems. Two objects can have the same mass, shape, colour, composition and be indistinguishable in all other intrinsic properties; but at least

* Objects, the unalterable, and the subsistent are one and the same. Objects are what is unalterable and subsistent; their configuration is what is changing and unstable.

they will differ in their position, or their velocity, or their orientation. The state pinpoints the *individuality* of a physical system,* and allows us to distinguish it from exact copies of itself. Therefore, the state also describes the relation of an object or a system with respect to its environment. Or in short: *the state describes all aspects of a system that depend on the observer.* These properties are not boring – just ponder this: does the universe have a state?

Challenge 10 n

Describing nature as a collection of permanent entities and changing states is the starting point of the study of motion. The various aspects of objects and of their states are called *observables*. All these rough, preliminary definitions will be refined step by step in the following. Using the terms just introduced, we can say that *motion is the change of state of objects.**

States are required for the description of motion. In order to proceed and to achieve a *complete* description of motion, we thus need a complete description of objects and a complete description of their possible states. The first approach, called Galilean physics, consists in specifying our *everyday* environment as precisely as possible.

Curiosities and fun challenges about motion

Motion is not always a simple topic.**

Challenge 11 n

- Is the motion of a ghost an example of motion?

- A man climbs a mountain from 9 a.m. to 1 p.m. He sleeps on the top and comes down the next day, taking again from 9 a.m. to 1 p.m. for the descent. Is there a place on the path that he passes at the same time on the two days?

Challenge 12 n

- Can something stop moving? If yes: how would you show it? If not: does this mean that nature is infinite?

Challenge 13 n

- Can the universe move?

Challenge 14 n

- To talk about precision with precision, we need to measure it. How would you do that?

Challenge 15 n

- Would we observe motion if we had no memory?

Challenge 16 n

- What is the lowest speed you have observed? Is there a lowest speed in nature?

Challenge 17 n

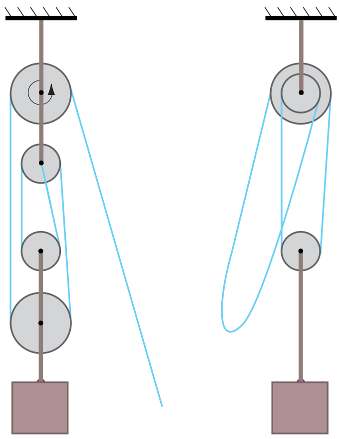


Figure 8 A block and tackle and a differential pulley

- According to legend, Sessa ben Zahir, the Indian inventor of the game of chess, demanded from King Shirham the following reward for his invention: he wanted one grain of rice for the first

* A *physical system* is a localized entity of investigation. In the classification of Table 2, the term 'physical system' is the same as 'object' or 'physical body'. Images are usually not counted as physical systems. Are holes physical systems?

Challenge 9 ny

* The exact separation between those aspects belonging to the object and those belonging to the state depends on the precision of observation. For example, the length of a piece of wood is not permanent; it shrinks and bends with time, due to processes at the molecular level. To be precise, the length of a piece of wood is not an aspect of the object, but an aspect of its state. Precise observations thus *shift* the distinction between the object and its state; the distinction itself does not disappear – at least for quite while.

** Sections entitled 'curiosities' are collections of topics and problems that allow one to check and to expand the usage of concepts introduced before.

square, two for the second, four for the third, eight for the fourth, and so on. How much time would all the rice fields of the world take to produce the necessary rice? Challenge 18 n

▪ When moving a burning candle, the flame lags behind. How does the flame behave if the candle is inside a glass, still burning, and the glass is accelerated? Challenge 19 n

▪ A good way to make money is to build motion detectors. A motion detector is a small box with a few wires. The box produces an electrical signal whenever the box moves. What types of motion detectors can you imagine? How cheap can you make such a box? How precise? Challenge 20 d

▪ A perfectly frictionless and spherical ball lies near the edge of a perfectly flat and horizontal table. What happens? In what time scale? Challenge 21 d

▪ You step into a closed box without windows. The box is moved by outside forces unknown to you. Can you determine how you move from inside the box? Challenge 22 n

▪ What is the length of rope one has to pull in order to lift a mass by a height h with a block and tackle with four wheels, as shown in Figure 8? Challenge 23 n

▪ When a block is rolled over the floor over a set of cylinders, how are the speed of the block and that of the cylinders related? Challenge 24 n

Ref. 12 ▪ Do you dislike formulae? If you do, use the following three-minute method to change the situation. It is worth trying it, as it will make you enjoy this book much more. Life is short; as much of it as possible, like reading this text, should be a pleasure. Challenge 25 n

1 - Close your eyes and recall an experience that was *absolutely marvellous*, a situation when you felt excited, curious and positive.

2 - Open your eyes for a second or two and look at page 292 – or any other page that contains many formulae.

3 - Then close your eyes again and return to your marvellous experience.

4 - Repeat the observation of the formulae and the visualization of your memory – steps 2 and 3 – three more times.

Then leave the memory, look around yourself to get back into the here and now, and test yourself. Have again a look at page 292. How do you feel about formulae now?

▪ In the sixteenth century, Niccolò Tartaglia* proposed the following problem. Three young couples want to cross a river. Only a small boat that can carry two people is available. The men are extremely jealous, and would never leave their brides alone with another man. How many journeys across the river are necessary? Challenge 26 n

3. Galilean physics – motion in everyday life

> Physic ist wahrlich das eigentliche Studium des Menschen.**
> Georg Christoph Lichtenberg

The simplest description of motion is the one we all, like cats or monkeys, use unconsciously in everyday life: *only one thing can be at a given spot at a given time*. This general description can be separated into three assumptions: matter is *impenetrable* and *moves*, time is made of *instants*, and space is made of *points*. Without these three assumptions (do you agree with them?) it is not possible to define velocity in everyday life. This description Challenge 27 n

* Niccolò Fontana Tartaglia (1499–1557), important Venetian mathematician.

** 'Physics truly is the proper study of man.' Georg Christoph Lichtenberg (1742–1799) was an important physicist and essayist.

of nature is called *Galilean* or *Newtonian physics*.

Galileo Galilei (1564–1642), Tuscan professor of mathematics, was a founder of modern physics and famous for advocating the importance of observations as checks of statements about nature. By requiring and performing these checks throughout his life, he was led to continuously increase the accuracy in the description of motion. For example, Galileo studied motion by measuring change of position with a self-constructed stopwatch. His approach changed the speculative description of ancient Greece into the experimental physics of Renaissance Italy.*

Galileo Galilei

The English alchemist, occultist, theologian, physicist and politician Isaac Newton (1643–1727) was one of the first to pursue with vigour the idea that different types of motion have the same properties, and made important steps in constructing the concepts necessary to demonstrate this idea.**

What is velocity?

> There is nothing else like it.
>
> Jochen Rindt***

Velocity fascinates. To physicists, not only car races are interesting, but any moving entity is. Therefore they first of all measure as many examples as possible. A selection is given in Table 4.

Everyday life teaches us a lot about motion: objects can overtake each other, and they can move in different directions. We also observe that velocities can be added or changed smoothly. The precise list of these properties, as given in Table 3, is summarized by mathematicians with a special term; they say that velocities form a *Euclidean vector space*.****
Page 66 More details about this strange term will be given shortly. For now we just note that in describing nature, mathematical concepts offer the most accurate vehicle.

When velocity is assumed to be an Euclidean vector, it is called *Galilean* velocity. Velocity is a profound concept. For example, velocity does not need space and time measurements to be defined first. Are you able to find a means to measure velocities without Ref. 22

* The best and most informative book on the life of Galileo and his times is by Pietro Redondi (see the footnote on page 203). Galileo was born in the year the pencil was invented. Before his time, it was impossible to do paper and pencil calculations. For the curious, the http://www.mpiwg-berlin.mpg.de website allows you to read an original manuscript by Galileo.

** Newton was born a year after Galileo died. Newton's other hobby, as master of the mint, was to supervise personally the hanging of counterfeiters. About Newton's infatuation with alchemy, see the books by Dobbs. Ref. 20 Among others, Newton believed himself to be chosen by god; he took his Latin name, Isaacus Neuutonus, and formed the anagram Jeova sanctus unus. About Newton and his importance for classical mechanics, see the text by Clifford Truesdell. Ref. 21

*** Jochen Rindt (1942–1970), famous Austrian Formula One racing car driver, speaking about speed.

**** It is named after Euclid, or Eukleides, the great Greek mathematician who lived in Alexandria around 300 BCE. Euclid wrote a monumental treatise of geometry, the Στοιχεῖα or *Elements*, which is one of the milestones of human thought. The text presents the whole knowledge on geometry of that time. For the first time, Euclid introduces two approaches that are now in common use: all statements are deduced from a small number of basic 'axioms' and for every statement a 'proof' is given. The book, still in print today, has been the reference geometry text for over 2000 years. On the web, it can be found at http://aleph0.clarku.edu/~djoyce/java/elements/elements.html.

Table 3 Properties of everyday – or Galilean – velocity

Velocities can	Physical property	Mathematical name	Definition
Be distinguished	distinguishability	element of set	Page 599
Point somewhere	direction	vector space, dimensionality	Page 66
Be added	additivity	vector space	Page 66
Change gradually	continuum	real vector space	Page 66, Page 1116
Have defined angles	direction	Euclidean vector space	Page 66
Be compared	measurability	metricity	Page 1108
Exceed any limit	infinity	unboundedness	Page 600

Challenge 28 d measuring space and time? If so, you probably want to continue reading on page 249, jumping 2000 years of enquiries. If you cannot do so, consider this: whenever we measure a quantity we assume that everybody is able to do so, and that everybody will get the same result. In other words, we take *measurement* to be a comparison with a standard. We thus implicitly assume that such a standard exists, i.e. that an example of a 'perfect' velocity can be found. Historically, the study of motion did not investigate this question first, because for many centuries nobody could find such a standard velocity. You are thus in good company.

Velocity is a profound subject for a second reason: we will discover that all properties of Table 3 are only approximate; *none* is actually correct. Improved experiments will uncover limits in every property of Galilean velocity. The failure of the last two properties will lead us to special and general relativity, the failure of the middle two to quantum theory and the failure of the first two properties to the unified description of nature. But for now, we'll stick with Galilean velocity, and continue with another Galilean concept derived from it: time.

> Without the concepts *place*, *void* and *time*, change cannot be. [...] It is therefore clear [...] that their investigation has to be carried out, by studying each of them separately.
>
> Aristotle* *Physics*, Book III, part 1.

What is time?

> Time does not exist in itself, but only through the perceived objects, from which the concepts of past, of present and of future ensue.
>
> Lucrece,** *De rerum natura*, lib. 1, v. 460 ss.

In their first years of life, children spend a lot of time throwing objects around. The term 'object' is a Latin word meaning 'that which has been thrown in front.' Developmental Ref. 15 psychology has shown experimentally that from this very experience children extract

* Aristotle (384/3–322), Greek philosopher and scientist.

** Lucretius Carus (*c.* 95 to *c.* 55 BCE), Roman scholar and poet.

Table 4 Some measured velocity values

Observation	Velocity
Stalagmite growth	0.3 pm/s
Can you find something slower?	Challenge 29 n
Lichen growth	down to 7 pm/s
Typical motion of continents	10 mm/a = 0.3 nm/s
Human growth during childhood, hair growth	4 nm/s
Tree growth	up to 30 nm/s
Electron drift in metal wire	1 µm/s
Sperm motion	60 to 160 µm/s
Speed of light at Sun's centre	0.1 mm/s
Ketchup motion	1 mm/s
Slowest speed of light measured in matter on Earth	0.3 m/s Ref. 23
Speed of snowflakes	0.5 m/s to 1.5 m/s
Signal speed in human nerve cells	0.5 m/s to 120 m/s Ref. 24
Wind speed at 1 Beaufort (light air)	below 1.5 m/s
Speed of rain drops, depending on radius	2 m/s to 8 m/s
Fastest swimming fish, sailfish (*Istiophorus platypterus*)	22 m/s
Fastest running animal, cheetah (*Acinonyx jubatus*)	30 m/s
Wind speed at 12 Beaufort (hurricane)	above 33 m/s
Speed of air in throat when sneezing	42 m/s
Fastest measured throw: cricket bowl	45 m/s
Freely falling human	50 to 90 m/s
Fastest bird, diving *Falco peregrinus*	60 m/s
Fastest badminton serve	70 m/s
Average speed of oxygen molecule in air at room temperature	280 m/s
Sound speed in dry air at sea level and standard temperature	330 m/s
Cracking whip's end	750 m/s
Speed of a rifle bullet	3 km/s
Speed of crack propagation in breaking silicon	5 km/s
Highest macroscopic speed achieved by man – the Voyager satellite	14 km/s
Average (and peak) speed of lightning tip	600 km/s (50 000 km/s)
Speed of Earth through universe	370 km/s
Highest macroscopic speed measured in our galaxy	$0.97 \cdot 10^8$ m/s Ref. 25
Speed of electrons inside a colour tv	$1 \cdot 10^8$ m/s
Speed of radio messages in space	299 972 458 m/s
Highest ever measured group velocity of light	$10 \cdot 10^8$ m/s
Speed of light spot from a light tower when passing over the Moon	$2 \cdot 10^9$ m/s
Highest proper velocity ever achieved for electrons by man	$7 \cdot 10^{13}$ m/s
Highest possible velocity for a light spot or shadow	infinite

the concepts of time and space. Adult physicists do the same when studying motion at university.

When we throw a stone through the air, we can define a *sequence* of observations. Our memory and our senses give us this ability. The sense of hearing registers the various sounds during the rise, the fall and the landing of the stone. Our eyes track the location of the stone from one point to the next. All observations have their place in a sequence, with some observations preceding them, some observations simultaneous to them, and still others succeeding them. We say that observations are perceived to happen at various *instants* and we call the sequence of all instants *time*.

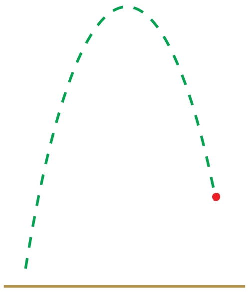


Figure 9 A typical path followed by a stone thrown through the air

An observation that is considered the smallest part of a sequence, i.e. not itself a sequence, is called an *event*. Events are central to the definition of time; in particular, starting or stopping a stopwatch are events. (But do events really exist? Keep this question in the back of your head as we move on.) Challenge 30 n

Sequential phenomena have an additional property known as stretch, extension or duration. Some measured valued are given in Table 5.* *Duration* expresses the idea that sequences *take* time. We say that a sequence takes time to express that other sequences can take place in parallel with it.

How exactly is the concept of time, including sequence and duration, deduced from observations? Many people have looked into this question: astronomers, physicists, watchmakers, psychologists and philosophers. All find that *time is deduced by comparing motions*. Children, beginning at a very young age, develop the concept of 'time' from the comparison of motions in their surroundings. Grown-ups take as a standard the motion of the Sun and call the resulting type of time *local time*. From the Moon they deduce a *lunar calendar*. If they take a particular village clock on a European island they call it the *universal time coordinate* (UTC), once known as 'Greenwich mean time.'** Astronomers use the movements of the stars and call the result *ephemeris time*. An observer who uses his personal watch calls the reading his *proper time*; it is often used in the theory of relativity. Ref. 15

Not every movement is a good standard for time. In the year 2000 an Earth rotation does not take 86 400 seconds any more, as it did in the year 1900, but 86 400.002 seconds. Can you deduce in which year your birthday will have shifted by a whole day? Page 1067 Challenge 32 n

All methods for the definition of time are thus based on comparisons of motions. In order to make the concept as precise and as useful as possible, a *standard* reference motion is chosen, and with it a standard sequence and a standard duration is defined. The device that performs this task is called a *clock*. We can thus answer the question of the section title: *time is what we read from a clock*. Note that all definitions of time used in the various branches of physics are equivalent to this one; no 'deeper' or more fundamental definition

* A year is abbreviated a (Latin 'annus').

** Official UTC time is used to determine power grid phase, phone companies' bit streams and the signal to the GPS system used by many navigation systems around the world, especially in ships, aeroplanes and lorries. For more information, see the http://www.gpsworld.com web site. The time-keeping infrastructure is also important for other parts of the modern economy as well. Can you spot the most important ones? Challenge 31 n

Table 5 Selected time measurements

Observation	Time
Shortest measurable time	10^{-44} s
Shortest time ever measured	10^{-23} s
Time for light to cross an atom	10^{-18} s
Period of caesium ground state hyperfine transition	108.782 775 707 78 ps
Beat of wings of fruit fly	1 ms
Period of pulsar (rotating neutron star) PSR 1913+16	0.059 029 995 271(2) s
Human 'instant'	20 ms
Shortest lifetime of living being	0.3 d
Average length of day 400 million years ago	79 200 s
Average length of day today	86 400.002(1) s
From birth to your 1000 million seconds anniversary	31.7 a
Age of oldest living tree	4600 a
Use of human language	$2 \cdot 10^5$ a
Age of Himalayas	35 to $55 \cdot 10^6$ a
Age of Earth	$4.6 \cdot 10^9$ a
Age of oldest stars	$14 \cdot 10^9$ a
Age of most protons in your body	$14 \cdot 10^9$ a
Lifetime of tantalum nucleus ^{180}Ta	10^{15} a
Lifetime of bismuth ^{209}Bi nucleus	$1.9(2) \cdot 10^{19}$ a

is possible.* Note that the word 'moment' is indeed derived from the word 'movement'. Language follows physics in this case. Astonishingly, the definition of time just given is final; it will never be changed, not even at the top of Motion Mountain. This is surprising at first sight, because many books have been written on the nature of time. Instead, they should investigate the nature of motion! But this is the aim of our walk anyhow. We are thus set to discover all the secrets of time as a side result of our adventure. Every clock reminds us that in order to understand time, we need to understand motion.

A *clock* is a moving system whose position can be read out. Of course, a *precise* clock is a system moving as regularly as possible, with as little outside disturbance as possible. Is there a perfect clock in nature? Do clocks exist at all? We will continue to study these questions throughout this work and eventually reach a surprising conclusion. At this point, however, we state a simple intermediate result: since clocks do exist, somehow there is in nature an intrinsic, natural and *ideal* way to measure time. Can you see it? Challenge 33 n

Time is not only an aspect of observations, it is also a facet of personal experience. Even in our innermost private life, in our thoughts, feelings and dreams, we experience sequences and durations. Children learn to relate this internal experience of time with external observations, and to make use of the sequential property of events in their actions.

* The oldest clocks are sundials. The science of making them is called *gnomonics*. An excellent and complete introduction into this somewhat strange world can be found at the http://www.sundials.co.uk website.

Table 6 Properties of Galilean time

Instants of time	Physical property	Mathematical name	Definition
Can be distinguished	distinguishability	element of set	Page 599
Can be put in order	sequence	order	Page 1116
Define duration	measurability	metricity	Page 1108
Can have vanishing duration	continuity	denseness, completeness	Page 1116
Allow durations to be added	additivity	metricity	Page 1108
Don't harbour surprises	translation invariance	homogeneity	Page 141
Don't end	infinity	unboundedness	Page 600
Can be defined for all observers	absoluteness	uniqueness	

Studies of the origin of psychological time show that it coincides – apart from its lack of accuracy – with clock time.* Every living human necessarily uses in his daily life the concept of time as a combination of sequence and duration; this fact has been checked Ref. 27 in numerous investigations. For example, the term 'when' exists in all human languages.

Time is a concept *necessary* to distinguish among observations. In any sequence, we observe that events succeed each other smoothly, apparently without end. In this context, 'smoothly' means that observations not too distant tend to be not too different. Yet between two instants, as close as we can observe them, there is always room for other events. Durations, or *time intervals*, measured by different people with different clocks agree in everyday life; moreover, all observers agree on the order in a sequence of events. Time is thus unique.

The mentioned properties of everyday time, listed in Table 6, correspond to the precise version of our everyday experience of time. It is called *Galilean time;* all the properties can be expressed simultaneously by describing time with real numbers. In fact, real numbers have been constructed to have exactly the same properties as Galilean time has, as Page 608 explained in the Intermezzo. Every instant of time can be described by a real number, often abbreviated t, and the duration of a sequence of events is given by the difference between the values for the final and the starting event.

When Galileo studied motion in the seventeenth century, there were no stopwatches yet. He thus had to build one himself, in order to measure times in the range between a Challenge 34 n fraction and a few seconds. Can you guess how he did it?

We will have quite some fun with Galilean time in the first two chapters. However, hundreds of years of close scrutiny have shown that *every single* property of time just listed is approximate, and none is strictly correct. This story is told in the subsequent chapters.

Page 774 * The brain contains numerous clocks. The most precise clock for short time intervals, the internal interval timer, is more accurate than often imagined, especially when trained. For time periods between a few tenths Ref. 26 of a second, as necessary for music, and a few minutes, humans can achieve accuracies of a few per cent.

Why do clocks go clockwise?

Challenge 35 n

> What time is it at the North Pole now?

All rotational motions in our society, such as athletic races, horse, bicycle or ice skating races, turn anticlockwise. Likewise, every supermarket leads its guests anticlockwise through the hall. Mathematicians call this the positive rotation sense. Why? Most people are right-handed, and the right hand has more freedom at the outside of a circle. Therefore thousands of years ago chariot races in stadia went anticlockwise. As a result, all races continue to do so to this day. That is why runners do it anticlockwise. For the same reason, helical stairs in castles are built in such a way that defending right-handers, usually from above, have that hand on the outside.

On the other hand, the clock imitates the shadow of sundials; obviously, this is true on the northern hemisphere only, and only for sundials on the ground, which were the most common ones. (The old trick to determine south by pointing the hour hand of an horizontal watch to the Sun and halving the angle between it and the direction of 12 o'clock does not work on the southern hemisphere.) So every clock implicitly continues to tell on which hemisphere it was invented. In addition, it also tells that sundials on walls came in use much later than those on the floor.

Does time flow?

> Wir können keinen Vorgang mit dem 'Ablauf der Zeit' vergleichen – diesen gibt es nicht –, sondern nur mit einem anderen Vorgang (etwa dem Gang des Chronometers).*
>
> Ludwig Wittgenstein, *Tractatus*, 6.3611

The expression 'the flow of time' is often used to convey that in nature change follows after change, in a steady and continuous manner. But though the hands of a clock 'flow', time itself does not. Time is a concept introduced specially to describe the flow of events around us; it does not itself flow, it *describes* flow. Time does not advance. Time is neither linear nor cyclic. The idea that time flows is as hindering to understanding nature as is the idea that mirrors exchange right and left. Page 512

The misleading use of the expression 'flow of time', propagated first by some Greek thinkers and then again by Newton, continues. Aristotle (384/3–322 BCE), careful to think logically, pointed out its misconception, and many did so after him. Nevertheless, expressions such as 'time reversal', the 'irreversibility of time', and the much-abused 'time's arrow' are still common. Just read a popular science magazine chosen at random. The fact is: time cannot be reversed, only motion can, or more precisely, only velocities of objects; time has no arrow, only motion has; it is not the flow of time that humans are unable to stop, but the motion of all the objects in nature. Incredibly, there are even books written by respected physicists which study different types of 'time's arrows' and compare them to each other. Predictably, no tangible or new result is extracted. Time does not flow. (Ref. 28; Challenge 36 e; Ref. 29)

In the same manner, colloquial expressions such as 'the start (or end) of time' should be avoided. A motion expert translates them straight away into 'the start (or end) of motion'.

* We cannot compare a process with 'the passage of time' – there is no such thing – but only with another process (such as the working of a chronometer).

What is space?

> The introduction of numbers as coordinates […] is an act of violence […].
>
> Hermann Weyl, *Philosophie der Mathematik und Naturwissenschaft*.*

Whenever we distinguish two objects from each other, such as two stars, we first of all distinguish their positions. Distinguishing positions is the main ability of our sense of sight. Position is therefore an important aspect of the physical state of an object. Positions are taken by only one object at a time. They are limited. The set of all available positions, called *(physical) space*, acts as both a container and a background.

Closely related to space and position is *size*, the set of positions an objects occupies. Small objects occupy only subsets of the positions occupied by large ones. We will discuss size shortly.

How do we deduce space from observations? During childhood, humans (and most higher animals) learn to bring together the various perceptions of space, namely the visual, the tactile, the auditory, the kinesthetic, the vestibular etc., into one coherent set of experiences and description. The result of this learning process is a certain 'image' of space in the brain. Indeed, the question 'where?' can be asked and answered in all languages of the world. Being more precise, adults derive space from distance measurements. The concepts of length, area, volume, angle and solid angle are all deduced with their help. Geometers, surveyors, architects, astronomers, carpet salesmen and producers of metre sticks base their trade on distance measurements. Space is a concept formed to summarize all the distance relations between objects for a precise description of observations.

Metre sticks work well only if they are straight. But when humans lived in the jungle, there were no straight objects around them. No straight rulers, no straight tools, nothing. Today, a cityscape is essentially a collection of straight lines. Can you describe how humans achieved this?

Challenge 37 n

Once humans came out of the jungle with their newly built metre sticks, they collected a wealth of results. The main ones are listed in Table 7; they are easily confirmed by personal experience. Objects can take positions in an apparently *continuous* manner: there indeed are more positions than can be counted.** Size is captured by defining the distance between various positions, called *length*, or by using the field of view an object takes when touched, called its *surface*. Length and surface can be measured with the help of a metre stick. Selected measurement results are given in Table 8. The length of objects is independent of the person measuring it, of the position of the objects and of their orientation. In daily life the sum of angles in any triangle is equal to two right angles. There are no limits in space.

Experience shows us that space has three dimensions; we can define sequences of positions in precisely three independent ways. Indeed, the inner ear of (practically) all vertebrates has three semicircular canals that sense the body's position in the three dimen-

* Hermann Weyl (1885–1955) was one of the most important mathematicians of his time, as well as an important theoretical physicist. He was one of the last universalists in both fields, a contributor to quantum theory and relativity, father of the term 'gauge' theory, and author of many popular texts.

** For a definition of uncountability, see page 602.

Table 7 Properties of Galilean space

Points	Physical property	Mathematical name	Definition
Can be distinguished	distinguishability	element of set	Page 599
Can be lined up if on one line	sequence	order	Page 1116
Can form shapes	shape	topology	Page 1116
Lie along three independent directions	possibility of knots	3-dimensionality	Page 1107
Can have vanishing distance	continuity	denseness, completeness	Page 1116
Define distances	measurability	metricity	Page 1108
Allow adding translations	additivity	metricity	Page 1108
Define angles	scalar product	Euclidean space	Page 66
Don't harbour surprises	translation invariance	homogeneity	
Can beat any limit	infinity	unboundedness	Page 600
Defined for all observers	absoluteness	uniqueness	Page 50

sions of space, as shown in Figure 10.* Similarly, each human eye is moved by three pairs of muscles. (Why three?) Another proof that space has three dimensions is provided by shoelaces: if space had more than three dimensions, shoelaces would not be useful, because knots exist only in three-dimensional space. But why does space have three dimensions? This is probably the most difficult question of physics; it will be answered only in the very last part of our walk.

Challenge 38 n

It is often said that thinking in four dimensions is impossible. That is wrong. Just try. For example, can you confirm that in four dimensions knots are impossible?

Challenge 39 n

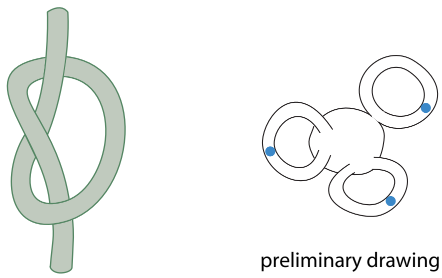


Figure 10 Two proofs of the three-dimensionality of space: a knot and the inner ear of a mammal

Like time intervals, length intervals can be described most precisely with the help of *real numbers*. In order to simplify communication, standard *units* are used, so that everybody uses the same numbers for the same length. Units allow us to explore the general properties of *Galilean space* experimentally: space, the container of objects, is continuous, three-dimensional, isotropic, homogeneous, infinite, Euclidean and unique or 'absolute'. In mathematics, a structure or mathematical concept with all the properties just mentioned is called a three-dimensional *Euclidean space*. Its elements, *(mathematical) points*, are described by three real parameters. They are usually written

* Note that saying that space has three dimensions *implies* that space is continuous; the Dutch mathematician and philosopher Luitzen E.J. Brouwer (b. 1881 Overschie, d. 1966 Blaricum) showed that dimensionality is only a useful concept for continuous sets.

as

$$(x, y, z) \tag{2}$$

and are called *coordinates*. They specify and order the location of a point in space. (For the precise definition of Euclidean spaces, see page 66.)

What is described here in just half a page actually took 2000 years to be worked out, mainly because the concepts of 'real number' and 'coordinate' had to be discovered first. The first person to describe points of space in this way was the famous mathematician and philosopher René Descartes*, after whom the coordinates of expression (2) are named *Cartesian*.

Like time, space is a *necessary* concept to describe the world. Indeed, space is automatically introduced when we describe situations with many objects. For example, when many spheres lie on a billiard table, we cannot avoid using space to describe the relations among them. There is no way to avoid using spatial concepts when talking about nature.

René Descartes

Even though we need space to talk about nature, it is still interesting to ask why this is possible. For example, since length measurement methods do exist, there must be a *natural* or *ideal* way to measure distances, sizes and straightness. Can you find
Challenge 40 n it?

As in the case of time, each of the properties of space just listed has to be checked. And again, careful observations will show that each property is an approximation. In simpler and more drastic words, *all* of them are wrong. This confirms Weyl's statement at the beginning of this section. In fact, the story about the violence connected with the introduction of numbers is told by every forest in the world, and of course also by the one at the foot of Motion Mountain. To hear it, we need only listen carefully to what the trees have to tell.

Μέτρον ἄριστον.**

Cleobulus

Are space and time absolute or relative?

In everyday life, the concepts of Galilean space and time include two opposing aspects; the contrast has coloured every discussion for several centuries. On one hand, space and time express something invariant and permanent; they both act like big containers for all the objects and events found in nature. Seen this way, space and time have an existence of their own. In this sense one can say that they are fundamental or *absolute*. On the other hand, space and time are tools of description that allow us to talk about relations between objects. In this view, they do not have any meaning when separated from objects, and only

* René Descartes or Cartesius (1596–1650), French mathematician and philosopher, author of the famous statement 'je pense, donc je suis', which he translated into 'cogito ergo sum' – I think therefore I am. In his view this is the only statement one can be sure of.

** 'Measure is the best (thing).' Cleobulus (Κλεοβουλος) of Lindos, (*c.* 620–550 BCE) was another of the proverbial seven sages.

Table 8 Some measured distance values

Observation	Distance
Galaxy Compton wavelength	10^{-85} m (calculated only)
Planck length, the shortest measurable length	10^{-32} m
Proton diameter	1 fm
Electron Compton wavelength	2.426 310 215(18) pm
Hydrogen atom size	30 pm
Smallest eardrum oscillation detectable by human ear	50 pm
Wavelength of visible light	0.4 to 0.8 μm
Size of small bacterium	5 μm
Point: diameter of smallest object visible with naked eye	20 μm
Diameter of human hair (thin to thick)	30 to 80 μm
Total length of DNA in each human cell	2 m
Largest living thing, the fungus *Armillaria ostoyae*	3 km
Length of Earth's Equator	40 075 014.8(6) m
Total length of human nerve cells	$8 \cdot 10^5$ km
Average Sun's distance	149 597 870 691(30) m
Light year	9.5 Pm
Distance to typical star at night	10 Em
Size of galaxy	1 Zm
Distance to Andromeda galaxy	28 Zm
Most distant visible object	125 Ym

result from the relations between objects; they are derived, relational or *relative*. Which of these viewpoints do you prefer? The results of physics have alternately favoured one viewpoint over the other. We will repeat this alternation throughout our adventure, until we find the solution. And obviously, it will turn out to be a third option.

Challenge 41 e

Ref. 30

Size: why area exists, but volume does not

A central aspect of objects is their size. As a small child, before school age, every human learns how to use the properties of size and space in his actions. As adults seeking precision, the definition of *distance* as the difference between coordinates allows us to define *length* in a reliable way. It took hundreds of years to discover that this is *not* the case. Several investigations in physics and mathematics led to complications.

The physical issues started with an astonishingly simple question asked by Lewis Richardson:* how long is the western coastline of Britain?

Following the coastline on a map using an odometer, a device shown in the Figure 11, Richardson found that the length l of the coastline depends on the scale s (say 1/10 000 or 1/500 000) of the map used:

$$l = l_0 \, s^{0.25} \tag{3}$$

* Lewis Fray Richardson (1881–1953), English physicist and psychologist.

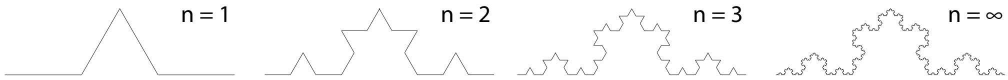


Figure 12 A fractal: a self-similar curve of infinite length (far right), and its construction

(Richardson found other numbers for other coasts.) The number l_0 is the length at scale 1:1. The main result is that the larger the map, the longer the coastline. What would happen if the scale of the map were increased even beyond the size of the original? The length would increase beyond all bounds. Can a coastline really have *infinite* length? Yes, it can. In fact, mathematicians have described many such curves; they are called *fractals*. An infinite number of them exists, and Figure 12 shows one example.* Can you construct another?

Challenge 42 e

Length has other strange properties. The Italian mathematician Giuseppe Vitali was the first to discover that it is possible to cut a line segment of length 1 into pieces that can be reassembled – merely by shifting them in direction of the segment – into a line segment of length 2. Are you able to find such a division using the hint that it is only possible using infinitely many pieces?

Challenge 43 d

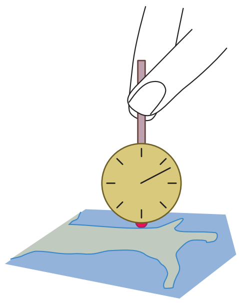


Figure 11 A curvemeter or odometer

In summary, length is well defined for lines that are straight or nicely curved, but not for intricate lines, or for lines made of infinitely many pieces. We therefore avoid fractals and other strangely shaped curves in the following, and we take special care when we talk about infinitely small segments. These are the central assumptions in the first two parts of this adventure, and we should never forget them. We will come back to these assumptions in the third part.

In fact, all these problems pale when compared to the following one. Commonly, area and volume are defined using length. You think that it is easy? You're wrong, as well as a victim of prejudices spread by schools around the world. To define area and volume with precision, their definitions must have two properties: the values must be *additive*, i.e. for finite and infinite sets of objects, the total area and volume have to be the sum of the areas and volumes of each element of the set; and they must be *rigid*, i.e. if one cuts an area or a volume into pieces and then rearranges them, the value remains the same. Do such concepts exist?

For areas in the plane, one proceeds in the following standard way: one defines the area A of a rectangle of sides a and b as $A = ab$; since any polygon can be rearranged into a rectangle with a finite number of straight cuts, one can then define an area value

Challenge 44 n

* Most of these curves are *self-similar*, i.e. they follow scaling laws similar to the above-mentioned. The term 'fractal' is due to the Polish mathematician Benoit Mandelbrot and refers to a strange property: in a certain sense, they have a non-integral number D of dimensions, despite being one-dimensional by construction. Mandelbrot saw that the non-integer dimension was related to the exponent e of Richardson by $D = 1 + e$, thus giving $D = 1.25$ in the example above.

Coastlines and other fractals are beautifully presented in Heinz-Otto Peitgen, Hartmut Jürgens & Dietmar Saupe, *Fractals for the Classroom*, Springer Verlag, 1992, pp. 232–245. It is available also in several other languages.

Page 162 for all polygons. Subsequently, one can define area for nicely curved shapes as the limit of the sum of infinitely many polygons. This method is called *integration*; it is introduced in detail in the section on physical action.

However, integration does not allow us to define area for arbitrarily bounded regions. Challenge 45 n (Can you imagine such a region?) For a complete definition, more sophisticated tools are needed. They were discovered in 1923 by the famous mathematician Stefan Banach.* He proved that one can indeed define an area for any set of points whatsoever, even if the border is not nicely curved but extremely complicated, such as the fractal curve just mentioned. Today this generalized concept of area, technically a 'finitely additive isometrically invariant measure,' is called a *Banach measure* in his honour. Mathematicians sum up this discussion by saying that since in two dimensions there is a Banach measure, there is a way to define the concept of area – an additive and rigid measure – for any set of points whatsoever.**

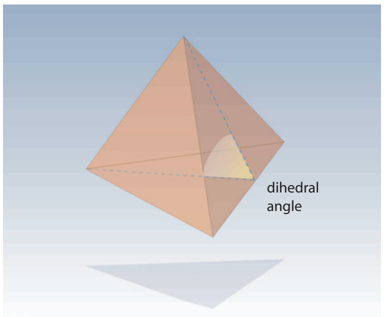


Figure 13 A polyhedron with one of its dihedral angles (© Luca Gastaldi)

What is the situation in *three* dimensions, i.e. for volume? We can start in the same way as for area, by defining the volume V of a rectangular polyhedron with sides a, b, c as $V = abc$. But then we encounter a first problem: a general polyhedron cannot be cut into a cube by straight cuts! The limitation was discovered in 1900 and 1902 by Max Dehn.*** He found that the possibility depends on the values of the edge angles, or dihedral angles, as the mathematicians call them. If one ascribes to every edge of a general polyhedron a number given by its length l times a special function $g(\alpha)$ of its dihedral angle α, then Dehn found that the sum of all the numbers for all the edges of a solid does not change under dissection, provided that the function fulfils $g(\alpha + \beta) = g(\alpha) + g(\beta)$ and $g(\pi) = 0$. An example of such a strange function g is the one assigning the value 0 to any rational multiple of π and the value 1 to a basis set of irrational multiples of π. The values for all other dihedral angles of the polyhedron can then be constructed by combination of rational multiples of these basis angles. Using this Challenge 46 n function, you may then deduce for yourself that a cube cannot be dissected into a regular tetrahedron because their respective Dehn invariants are different.****

Despite the problems with Dehn invariants, one *can* define a rigid and additive concept of volume for polyhedra, since for all of them and in general for all 'nicely curved' shapes, one can again use integration for the definition of their volume.

* Stefan Banach (Krakow, 1892–Lvov, 1945), Polish mathematician.

** Actually, this is true only for sets on the plane. For curved surfaces, such as the surface of a sphere, there are complications that will not be discussed here. In addition, the mentioned problems in the definition of length of fractals reappear also for area if the surface to be measured is not flat but full of hills and valleys. A typical example is the area of the human lung: depending on the level of details looked at, one finds area values from a few square metres up to over a hundred.

*** Max Dehn (1878–1952), German mathematician, student of David Hilbert.

**** This is also told in the beautiful book by M. Aigler & G.M. Ziegler, *Proofs from the Book*, Springer Verlag, 1999. The title is due to the famous habit of the great mathematician Paul Erdös to imagine that all beautiful mathematical proofs can be assembled in the 'book of proofs'.

Now let us consider general shapes and general cuts in three dimensions, not just the 'nice' ones mentioned so far. We then stumble on the famous *Banach–Tarski theorem* (or paradox). In 1924, Stefan Banach and Alfred Tarski* proved that it is possible to cut one sphere into five pieces that can be recombined to give two spheres, each the size of the original. This counter-intuitive result is the Banach–Tarski theorem. Even worse, another version of the theorem states: take any two sets not extending to infinity and containing a solid sphere each; then it is always possible to dissect one into the other with a *finite* number of cuts. In particular it is possible to dissect a pea into the Earth, or vice versa. Size does not count!** Volume is thus not a useful concept at all.

Ref. 31

The Banach–Tarski theorem raises two questions: First, can the result be applied to gold or bread? That would solve many problems. Second, can this blowing up be realized with empty space? In other words, are matter and empty space continuous? Both topics will be explored later in our walk; each issue will have its own, special consequences. For the moment, we eliminate this troubling issue by restricting our interest to smoothly curved shapes (and cutting knives). With this restriction, volumes of matter and of empty space do behave nicely: they are additive and rigid, and show no paradoxes. Indeed, the cuts required for the Banach–Tarski paradox are not smooth; it is not possible to perform them with an everyday knife, as they require (infinitely many) infinitely sharp bends performed with an infinitely sharp knife. Such a knife does not exist. Nevertheless, we keep in the back of our mind that the size of an object or of a piece of empty space is a tricky quantity – and that we need to be careful whenever we talk about it.

Challenge 47 n

What is straight?

When you see a solid object with a straight edge, it is a 99 %-safe bet that it is human made.* The contrast between the objects seen in a city – buildings, furniture, cars, electricity poles, boxes, books – and the objects seen in a forest – trees, plants, stones, clouds – is evident: in the forest nothing is straight or flat, in the city most objects are. How is it possible for humans to produce straight objects while there are none to be found in nature?

Any forest teaches us the origin of straightness; it presents tall tree trunks and rays of daylight entering from above through the leaves. For this reason we call a line *straight* if it touches either a plumb-line or a light ray along its whole length. In fact, the two definitions are equivalent. Can you confirm this? Can you find another definition? Obviously, we call a surface *flat* if for any chosen orientation and position it touches a plumb-line or a light ray along its whole extension.

Challenge 48 n

* Alfred Tarski (b. 1902 Warsaw, d. 1983 Berkeley), Polish mathematician.

** The proof of the result does not need much mathematics; it is explained beautifully by Ian Stewart in Paradox of the spheres, *New Scientist*, 14 January 1995, pp. 28–31. In 4 dimensions, the Banach–Tarski paradox exists as well, as it does in any higher dimension. More mathematical detail can be found in the beautiful book by Steve Wagon. Ref. 32

* The most common counterexamples are numerous crystalline minerals, where the straightness is related to the atomic structure. Another famous exception is the well-known Irish geological formation called Giant's Causeway. Other candidates which might come to mind, such as certain bacteria which have (almost) square or (almost) triangular shapes are not counterexamples, as the shapes are only approximate. Ref. 33

Figure 14 A photograph of the earth – seen from the direction of the sun

In summary, the concept of straightness – and thus also flatness – is defined with the help of bodies or radiation. In fact, all spatial concepts, like all temporal concepts, require motion for their definition.

A hollow Earth?

Space and straightness pose subtle challenges. Some strange people maintain that all humans live on the *inside* of a sphere; they (usually) call this the *hollow Earth theory*. They claim that the Moon, the Sun and the stars are all near the centre of the hollow sphere. They also explain that light follows curved paths in the sky and that when usual physicists talk about a distance r from the centre of the Earth, the real hollow Earth distance is $r_{\text{he}} = R^2_{\text{Earth}}/r$. Can you show that this model is wrong? Roman Sexl* used to ask this question to his students and fellow physicists. The answer is simple: if you think you have an argument to show that this view is wrong, you made a mistake! There is *no way* to show that such a view is wrong. It is possible to explain the horizon, the appearance of day and night as well as the satellite photographs of the round Earth, such as Figure 14. To explain what happened during the flight to the Moon is also fun. A coherent hollow Earth view is fully *equivalent* to the usual picture of an infinitely extended space. We will come back to this problem in the section on general relativity.

Challenge 49 n

Challenge 50 e

Page 458

* Roman Sexl, (1939–1986), important Austrian physicist, author of several influential textbooks on gravitation and relativity.

Figure 15 A model of the hollow Earth theory, showing how day and night appear

Curiosities and fun challenges about everyday space and time

Space and time lead to many thought-provoking questions.

▪ Imagine a black spot on a white surface. What is the colour of the line separating the spot from the background? This question is often called Peirce's puzzle. Challenge 51 ny

▪ Also bread is an (approximate) irregular fractal. The fractal dimension of bread is around 2.7. Try to measure it. Challenge 52 ny

▪ Motoring poses many mathematical problems. A central one is the following parking issue: what is the shortest gap d to the car parked in front necessary to leave a parking line without using reverse gear? (Assume that you know the geometry of your car, as shown in Figure 16, and its smallest outer turning radius R, which is known for every car.) Next question: what is the smallest gap required when you are allowed to manoeuvre back and forward as often as you like? Now a problem to which no solution seems to be available in the literature: How does the gap depend on the number n of times you use the reverse gear? (The author offers 50 Euro for the first well-explained solution sent to him.) Challenge 53 n Challenge 54 n Challenge 55 n

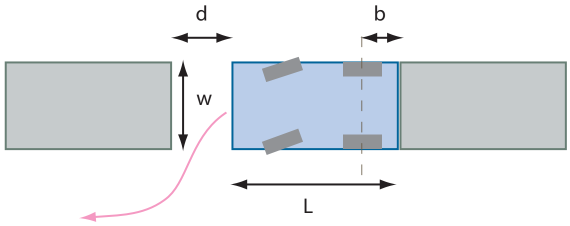


Figure 16 Leaving a parking space

▪ How often in 24 hours do the hour and minute hands of a clock lie on top of each

Table 9 The exponential notation: how to write small and large numbers

Number	Exponential notation	Number	Exponential notation
1	10^0		
0.1	10^{-1}	10	10^1
0.2	$2 \cdot 10^{-1}$	20	$2 \cdot 10^1$
0.324	$3.24 \cdot 10^{-1}$	32.4	$3.24 \cdot 10^1$
0.01	10^{-2}	100	10^2
0.001	10^{-3}	1000	10^3
0.000 1	10^{-4}	10 000	10^4
0.000 01	10^{-5} etc.	100 000	10^5 etc.

Challenge 56 n other? How often do all three hands lie on top of each other for clocks that also have a second hand?

▪ How many times in twelve hours can the two hands of a clock be *exchanged* with the result that the new situation shows a *valid* time? What happens for clocks having also a third hand for seconds? Challenge 57 n

▪ How many minutes does the Earth rotate in one minute? Challenge 58 n

▪ What is the highest speed achieved by throwing (with and without rackets)? What was the projectile used? Challenge 59 n

▪ A rope is put around the Earth, on the Equator, as tightly as possible. Then the rope is lengthened by 1 m. Can a mouse slip through? Challenge 60 n

▪ Jack was rowing his boat on a river. Below a bridge, he dropped his ball into the river. Jack continued to row in the same direction for 10 minutes after he dropped the ball. He then turned around and rowed back. When he reached the ball, the ball had floated 600 m from the bridge. How fast was the river flowing? Challenge 61 n

▪ Adam and Bert are brothers. Adam is 18 years old. Bert is twice as old as at the time when Adam was the age that Bert is now. How old is Bert?

▪ Scientists use a special way to write large and small numbers, explained in Table 9.

Ref. 34 In 1996 the smallest experimentally probed distance was 10^{-19} m, achieved between quarks at Fermilab. (To savour the distance value, write it down without exponent.) What does this measurement mean for the continuity of space? Challenge 62 n

▪ 'Where am I?' is a common question; 'when am I?' is never used, not even in other languages. Why? Challenge 63 n

▪ Is there a smallest time interval in nature? A smallest distance? Challenge 64 n

▪ Given that you know what straightness is, how would you characterize or define the curvature of a curved line using numbers? And that of a surface? Challenge 65 n

▪ What is the speed of your eyelid? Challenge 66 n

▪ The surface area of the human body is about 200 m^2. Can you say where this large number comes from? Challenge 67 e

▪ Fractals in three dimensions bear many surprises. Take a regular tetrahedron; then glue on every one of its triangular faces a smaller regular tetrahedron, so that the surface of the body is again made up of many equal regular triangles. Repeat the process, gluing

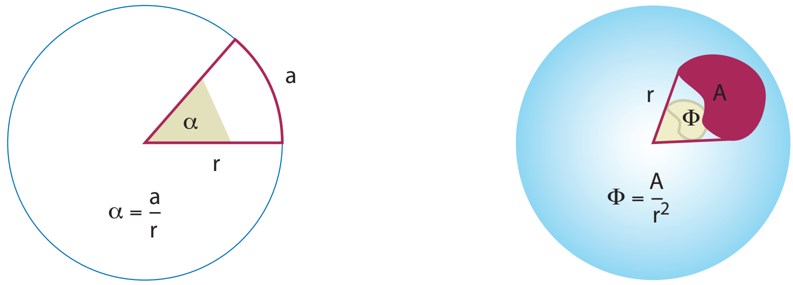


Figure 17 The definition of plane and solid angles

still smaller tetrahedrons to these new (more numerous) triangular surfaces. What is the shape of the final fractal, after an infinite number of steps? Challenge 68 n

▪ Zeno reflected on what happens to a moving object at a given instant of time. To discuss with him, you decide to build the fastest possible shutter for a photographic camera that you can imagine. You have all the money you want. What is the shortest shutter time you would achieve? Challenge 69 n

▪ Can you prove Pythagoras's theorem by geometrical means alone, without using coordinates? (There are more than 30 possibilities.) Challenge 70 n

▪ Why are most planets and moons (almost) spherical? Challenge 71 n

▪ A rubber band connects the tips of the two hands of a clock. What is the path followed by the middle point of that band? Challenge 72 n

▪ There are two important quantities connected to angles. As shown in Figure 17, what is usually called a *(plane) angle* is defined as the ratio between arc and radius lengths. A right angle is $\pi/2$ radian or $\pi/2$ rad or 90°.

The *solid angle* is the ratio between area and the square of the radius. An eighth of a sphere is $\pi/2$ or steradian $\pi/2$ sr. As a result, a small solid angle shaped like a cone and the angle of the cone tip are *different*. Can you find the relation? Challenge 73 n

▪ The definition of angle helps to determine the size of a firework display. Measure the time T between the moment that you see the rocket explode in the sky and the moment you hear the explosion, measure the (plane) angle α of the ball with your hand. The diameter D is

$$D \approx 6\,\mathrm{s}/^{\circ}\; T\, \alpha\,. \qquad (4)$$

Challenge 74 e Why? By the way, the angular distance between the knuckles of an extended fist are about 3°, 2° and 3°, the size of an extended hand 20°. For more about fireworks, see the http://cc.oulu.fi/~kempmp website.

▪ Measuring angular size with the eye only is tricky. For example, can you say whether the Moon is larger or smaller than the nail of your thumb at the end of your extended arm? Angular size is not an intuitive quantity; it requires measurement instruments. Challenge 75 e

A famous example, shown in Figure 18, illustrates the difficulty of estimating angles. Both the Sun and the Moon seem larger when they are on the horizon. In ancient times, Ptolemy explained this illusion by an unconscious apparent distance change induced by the human brain. In fact, the Moon is even *further away* from the observer when it is just above the horizon, and thus its image is *smaller* than a few hours earlier, when it was high

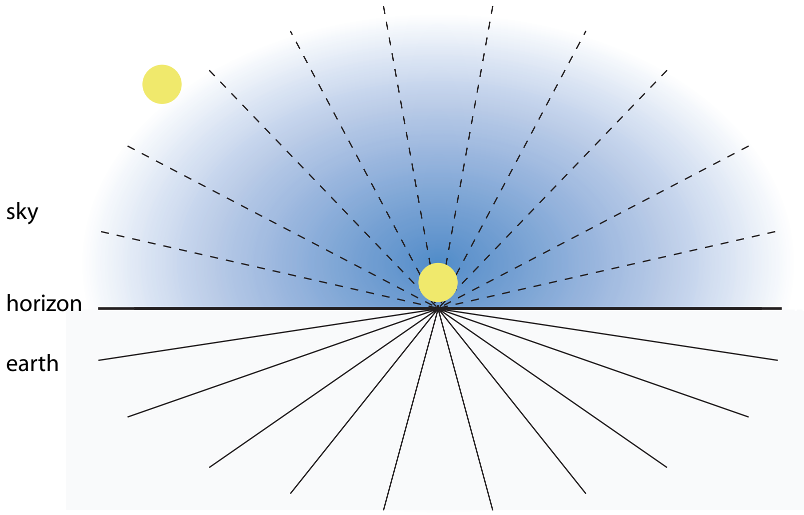


Figure 18 How the apparent size of the Moon and the Sun changes

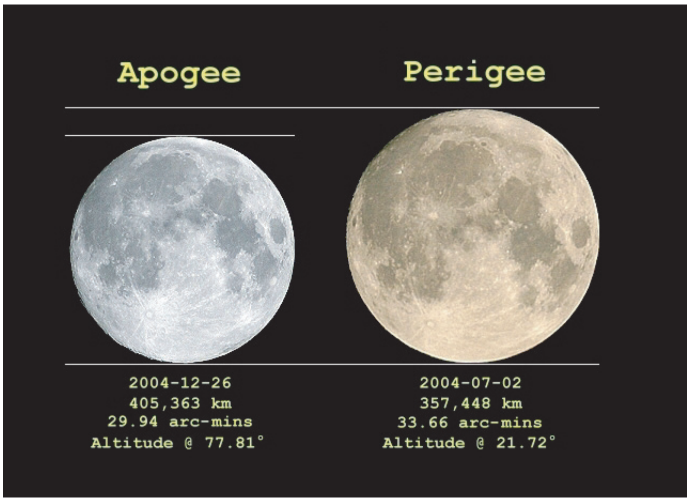


Figure 19 How the size of the Moon changes with distance (© Anthony Ayiomamitis)

Challenge 76 n in the sky. Can you confirm this?

In fact, the Moon's size changes much more due to another effect: the orbit of the Moon is elliptical. An example of the result is shown in Figure 19.

▪ Cylinders can be used to roll a flat object over the floor; they keep the object plane always at the same distance from the floor. What cross sections *other* than a circle allow you to realize the same feat? How many examples can you find? Challenge 77 n

▪ Galileo also made mistakes. In his famous book, the *Dialogues*, he says that the curve formed by a thin chain hanging between two nails is a parabola, i.e. the curve defined by $y = x^2$. That is not correct. What is the correct curve? You can observe the shape (approximately) in the shape of suspension bridges. Challenge 78 d

Figure 21 Anticrepuscular rays (© Peggy Peterson)

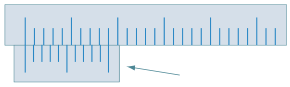


Figure 20 A vernier/nonius/clavius

▪ How does a *vernier* work? It is called *nonius* in other languages. The first name is derived from a French military engineer* who did not invent it, the second is a play of words on the latinized name of the Portuguese inventor of a more elaborate device** and the Latin word for 'nine'. In fact, the device as we know it today – shown in Figure 20 – was designed around 1600 by Christophonius Clavius,*** the same astronomer who made the studies that formed the basis of the Gregorian calendar reform of 1582. Are you able to design a vernier/nonius/clavius which instead of increasing the precision tenfold, does so by an arbitrary factor? Is there a limit to the attainable precision? Challenge 79 n

▪ Draw three circles, of different sizes, that touch each other. Now draw a fourth circle in the space between, touching the outer three. What simple relation do the inverse radii of the four circles obey? Challenge 80 n

▪ Take a tetrahedron OABC whose triangular sides OAB, OBC and OAC are rectangular in O. In other words, OA, OB and OC are all perpendicular to each other. In the tetrahedron, the areas of the triangles OAB, OBC and OAC are respectively 8, 4 and 1. What is the area of triangle ABC? Challenge 81 ny

▪ With two rulers, you can add and subtract numbers by lying them side by side. Are you able to design rulers that allow you to multiply and divide in the same manner? More elaborate devices using this principle were called *slide rules* and were the precursors of electronic calculators; they were in use all over the world until the 1970s. Challenge 82 n

▪ How many days would a year have if the Earth turned the other way with the same rotation frequency? Challenge 83 n

▪ Where is the Sun in the spectacular situation shown in Figure 21? Challenge 84 n

▪ Could a two-dimensional universe exist? Alexander Dewdney described such a universe in a book. Can you explain why a two-dimensional universe is impossible? Ref. 35 Challenge 85 ny

* Pierre Vernier (1580–1637), French military officer interested in cartography.
** Pedro Nuñes or Peter Nonnius (1502–1578), Portuguese mathematician and cartographer.
*** Christophonius Clavius or Schlüssel (1537–1612), Bavarian astronomer.

How to describe motion: kinematics

La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi agli occhi (io dico l'universo) ... Egli è scritto in lingua matematica.*

Galileo Galilei, *Il saggiatore* VI.

Experiments show that the properties of Galilean time and space are extracted from the environment by most higher animals and by young children. Later, when children learn to speak, they put these experiences into concepts, as was just done above. With the help of these concepts, grown-up children then say that *motion is change of position with time*. This description is illustrated by flipping rapidly the lower left corners starting at page 158. Each page simulates an instant of time, and the only change taking place during motion is the position of the object, represented by the dark spot. The other variations from one picture to the next, due to the imperfections of printing techniques, can be taken to simulate the inevitable measurement errors.

It is evident that calling 'motion' the change of position with time is *neither* an explanation *nor* a definition, since both the concepts of time and position are deduced from motion itself. It is only a *description* of motion. Still, the description is useful, because it allows for high precision, as we will find out exploring gravitation and electrodynamics. After all, precision is our guiding principle during this promenade. Therefore the detailed description of changes in position has a special name: it is called *kinematics*.

The set of all positions taken by an object over time forms a *path* or *trajectory*. The origin of this concept is evident when one watches fireworks* or again the previously mentioned flip movie in the lower left corners after page 158. With the description of space and time by real numbers, a trajectory can be described by specifying its three coordinates (x, y, z) – one for each dimension – as continuous functions of time t. (Functions are defined in detail on page 603.) This is usually written as $\mathbf{x} = \mathbf{x}(t) = (x(t), y(t), z(t))$. For example, observation shows that the height z of any thrown or falling stone changes as

$$z(t) = z_0 + v_0(t - t_0) - \frac{1}{2}g(t - t_o)^2 \qquad (5)$$

where t_0 is the time one starts the experiment, z_0 is the initial height, v_0 is the initial velocity in the vertical direction and $g = 9.8\,\mathrm{m/s^2}$ is a constant that is found to be the same, within about one part in 300, for all falling bodies on all points of the surface of the Earth. Where do the value $9.8\,\mathrm{m/s^2}$ and its slight variations come from? A preliminary answer will be given shortly, but the complete elucidation will occupy us during the larger part of this hike.

Ref. 36

Equation (5) allows us to determine the depth of a well, given the time a stone takes to reach its bottom. The equation also gives the speed v with which one hits the ground after jumping from a tree, namely $v = \sqrt{2gh}$. A height of 3 m yields a velocity of 27 km/h.

Challenge 86 n

* Science is written in this huge book that is continuously open before our eyes (I mean the universe) ... It is written in mathematical language.

* On the world of fireworks, see the frequently asked questions list of the usenet group rec.pyrotechnics, or search the web. A simple introduction is the article by J.A. Conkling, Pyrotechnics, *Scientific American* pp. 66–73, July 1990.

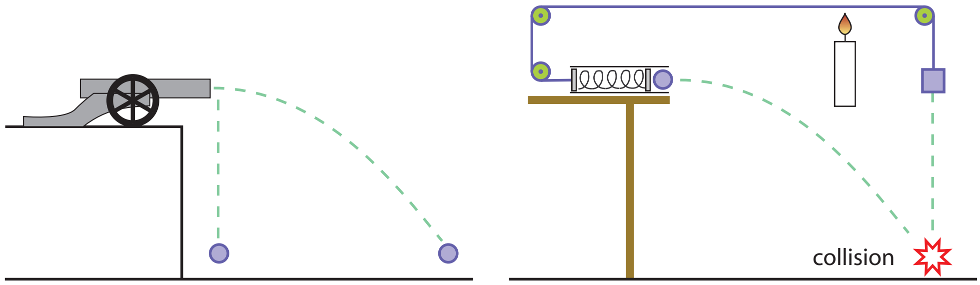


Figure 22 Two ways to test that the time of free fall does not depend on horizontal velocity

Challenge 87 n The velocity is thus proportional only to the square root of the height. Does this mean that strong fear of falling results from an overestimation of its actual effects?

Galileo was the first to state an important result about free fall: the motions in the horizontal and vertical directions are *independent*. He showed that the time of fall of a cannon ball shot exactly horizontally is *independent* of the strength of the gunpowder, as shown in Figure 22. Since many great thinkers did not agree with this statement even Ref. 37 after his death, in 1658 the Academia del Cimento even organized an experiment to check this assertion, by comparing the flying cannon ball to one that simply fell vertically. Can Challenge 88 n you imagine how they checked the simultaneity? Figure 22 also shows how to realize this check at home. Whatever the load of the cannon, the two bodies will always collide, thus proving the assertion.

In other words, a canon ball is not accelerated in the horizontal direction. Its horizontal motion is simply unchanging. By extending the description of equation (5) with the two expressions for the horizontal coordinates x and y, namely

$$\begin{aligned} x(t) &= x_0 + v_{x0}(t - t_0) \\ y(t) &= y_0 + v_{y0}(t - t_0) \ , \end{aligned} \tag{6}$$

a *complete* description for the path followed by thrown stones results. A path of this shape is called a *parabola*; it is shown in Figures 9, 22 and 23.* A parabola is also the shape used Challenge 89 n for light reflectors inside pocket lamps or car headlights. Can you show why?

The kinematic description of motion is useful for answering a whole range of questions:

▪ Numerous species of moth and butterfly caterpillars shoot away their frass – to speak Ref. 38 vulgarly: their shit – to prevent its smell from helping predators to locate them. Stanley Caveney and his team took photographs of this process. Figure 24 shows a caterpillar (yellow) of the skipper *Calpodes ethlius* inside a rolled up green leaf caught in the act. Given that the record distance observed is 1.5 m (though by another species, *Epargyreus* Challenge 90 n *clarus*), what is the ejection speed? How do caterpillars achieve it?

* Apart from the graphs shown in Figure 23, there is also the *configuration space* spanned by the coordinates of all particles of a system; only for a single particle it is equal to the real space. The phase space diagram is also called state space diagram.

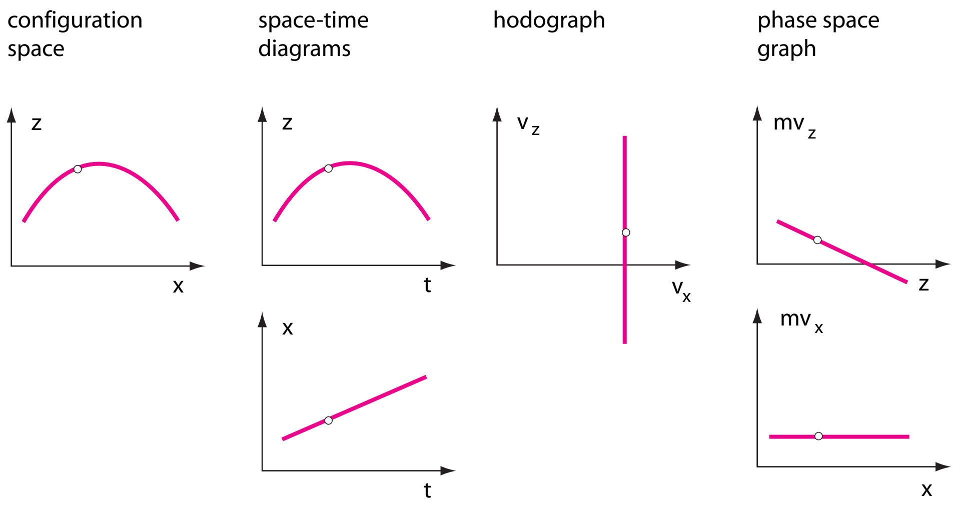


Figure 23 Various types of graphs describing the same flying stone

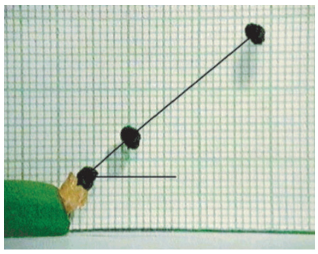


Figure 24 Three superimposed images of a frass pellet shot away by a caterpillar

▪ What is the horizontal distance one can reach with a stone, given the speed and the angle from the horizontal at which it is thrown?

Challenge 91 n

▪ How can the speed of falling rain be measured with an umbrella?

Challenge 92 n

▪ What is the maximum numbers of balls that could be juggled at the same time?

Challenge 93 n

▪ Finding an upper limit for the long jump is interesting. The running speed world record in 1997 was 12 m/s ≈ 43 km/h by Ben Johnson, and the women's record was 11 m/s ≈ 40 km/h. In fact, long jumpers never run much faster than about 9.5 m/s. How much extra jump distance could they achieve if they could run full speed? How could they achieve that? In addition, long jumpers take off at angles of about 20°, as they are not able to achieve a higher angle at the speed they are running. How much would they gain if they could achieve 45°?

Ref. 39 Ref. 40 Challenge 94 n

▪ Is it true that rain drops would kill if it weren't for the air resistance of the atmosphere? What about ice?

Challenge 95 n

▪ Are gun bullets falling back after being fired into the air dangerous?

Challenge 96 n

The last two questions arise because equation (5) does not hold in all cases. For example, leaves or potato crisps do not follow it. As Galileo already knew, this is a consequence of air resistance; we will discuss it shortly. In fact, even without air resistance, the path of a stone is not always a parabola; can you find such a situation?

Challenge 97 n

What is rest?

In the Galilean description of nature, motion and rest are opposites. In other words, a body is at rest when its position, i.e. its coordinates, do not change with time. In other words, (Galilean) *rest* is defined as

$$\mathbf{x}(t) = \text{const} . \tag{7}$$

Later we will see that this definition, contrary to first impressions, is not of much use and will have to be modified. The definition of rest implies that non-resting objects can be distinguished by comparing the rapidity of their displacement. One thus can define the *velocity* $\mathbf{v}$ of an object as the change of its position $\mathbf{x}$ with time t. This is usually written as

$$\mathbf{v} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} . \tag{8}$$

In this expression, valid for each coordinate separately, $\mathrm{d}/\mathrm{d}t$ means 'change with time'; one can thus say that velocity is the *derivative* of position with respect to time. The *speed* v is the name given to the magnitude of the velocity $\mathbf{v}$. Derivatives are written as fractions in order to remind the reader that they are derived from the idea of slope. The expression

$$\frac{\mathrm{d}y}{\mathrm{d}t} \quad \text{is meant as an abbreviation of} \quad \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} , \tag{9}$$

a shorthand for saying that the *derivative at a point* is the limit of the slopes in the neighbourhood of the point, as shown in Figure 25. This definition implies the working rules

Challenge 98 e

$$\frac{\mathrm{d}(y+z)}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\mathrm{d}z}{\mathrm{d}t} \quad , \quad \frac{\mathrm{d}(cy)}{\mathrm{d}t} = c\frac{\mathrm{d}y}{\mathrm{d}t} \quad , \quad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}^2 y}{\mathrm{d}t^2} \quad , \quad \frac{\mathrm{d}(yz)}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}t}z + y\frac{\mathrm{d}z}{\mathrm{d}t} \tag{10}$$

c being any number. This is all one ever needs to know about derivatives. The quantities $\mathrm{d}t$ and $\mathrm{d}y$, sometimes useful by themselves, are called *differentials*. These concepts are due to Gottfried Wilhelm Leibniz.* Derivatives lie at the basis of all calculations based on the continuity of space and time. Leibniz was the person who made it possible to describe and use velocity in physical formulae, and in particular, to use the idea of velocity at a given point in time or space for calculations.

* Gottfried Wilhelm Leibniz (b. 1646 Leipzig, d. 1716 Hannover), Saxon lawyer, physicist, mathematician, philosopher, diplomat and historian. He was one of the great minds of mankind; he invented the differential calculus (before Newton) and published many successful books in the various fields he explored, among them *De Arte Combinatoria*, *Hypothesis Physica Nova*, *Discours de métaphysique*, *Nouveaux essais sur l'entendement humain*, the *Théodicée* and the *Monadologia*.

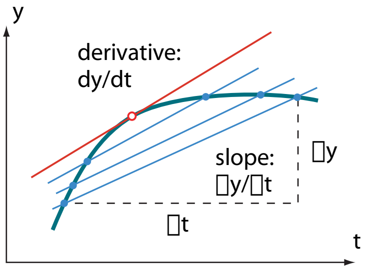


Figure 25 Derivatives

The definition of velocity assumes that it makes sense to take the limit $\Delta t \to 0$. In other words, it is assumed that infinitely small time intervals do exist in nature. The definition of velocity with derivatives is possible only because both space and time are described by sets which are *continuous*, or in mathematical language, *connected and complete*. In the rest of our walk we shall not forget that right from the beginning of classical physics, *infinities* are present in its description of nature. The infinitely small is part of our definition of velocity. Indeed, differential calculus can be defined as the study of infinity and its uses. We thus discover that the appearance of infinity does not automatically render a description impossible or imprecise. In order to remain precise, physicists use only the smallest two of the various possible types of infinities. Their precise definition and an overview of other types are introduced in the intermezzo following this chapter. Page 601

Gottfried Leibniz

The appearance of infinity in the usual description of motion was first criticized in his famous ironical arguments by Zeno of Elea (around 445 BCE), a disciple of Parmenides. In his so-called third argument, Zeno explains that since at every instant a given object occupies a part of space corresponding to its size, the notion of velocity at a given instant makes no sense; he provokingly concludes that therefore motion does not exist. Nowadays we would not call this an argument against the *existence* of motion, but against its usual *description*, in particular against the use of infinitely divisible space and time. (Do you agree?) Nevertheless, the description criticized by Zeno actually works quite well in everyday life. The reason is simple but deep: in daily life, changes are indeed continuous.

Ref. 41

Challenge 99 e

Large changes in nature are made up of many small changes. This property of nature is not obvious. For example, we note that we have tacitly assumed that the path of an object is not a fractal or some other badly behaved entity. In everyday life this is correct; in other domains of nature it is not. The doubts of Zeno will be partly rehabilitated later in our walk, and the more so the more we will proceed. The rehabilitation is only partial, as the solution will be different from what he ever dreamt of; on the other hand, the doubts on the idea of 'velocity at a point' will turn out to be well-founded. For the moment though, we have no choice: we continue with the basic assumption that in nature changes happen smoothly.

Page 923

Why is velocity necessary as a concept? Aiming for precision in the description of motion, we need to find the complete list of aspects necessary to specify the state of an object. The concept of velocity is obviously a member of this list. Continuing along the same lines, we call *acceleration* $\mathbf{a}$ of a body the change of velocity $\mathbf{v}$ with time, or

$$\mathbf{a} = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}t^2} . \tag{11}$$

Acceleration is what we feel when the Earth trembles, an aeroplane takes off, or a bicycle goes round a corner. More examples are given in Table 10. Like velocity, acceleration has both a magnitude and a direction, properties indicated by the use of **bold** letters for their abbreviations. *

Higher derivatives than acceleration can also be defined in the same manner. They add Ref. 42 little to the description of nature, because as we will show shortly neither they nor even acceleration itself are useful for the description of the state of motion of a system.

Objects and point particles

> Wenn ich den Gegenstand kenne, so kenne ich auch sämtliche Möglichkeiten seines Vorkommens in Sachverhalten.**
>
> Ludwig Wittgenstein, *Tractatus*, 2.0123

One aim of the study of motion is to find a complete and precise description of both states and objects. With the help of the concept of space, the description of objects can be refined considerably. In particular, one knows from experience that all objects seen in daily life Challenge 102 e have an important property: they can be divided into *parts*. Often this observation is expressed by saying that all objects, or bodies, have two properties. First, they are made out of *matter*,*** defined as that aspect of an object responsible for its impenetrability, i.e. the property preventing two objects from being in the same place. Secondly, bodies

* Such physical quantities are called *vectors*. In more precise, mathematical language, a vector is an element of a set, called *vector space*, in which the following properties hold for all vectors $\mathbf{a}$ and $\mathbf{b}$ and for all numbers c and d:

$$c(\mathbf{a}+\mathbf{b}) = c\mathbf{a}+c\mathbf{b} \quad , \quad (c+d)\mathbf{a} = c\mathbf{a}+d\mathbf{a} \quad , \quad (cd)\mathbf{a} = c(d\mathbf{a}) \quad \text{and} \quad 1\mathbf{a} = \mathbf{a} . \tag{12}$$

Another example of vector space is the set of all *positions* of an object. Does the set of all rotations form a Challenge 100 n vector space? All vector spaces allow the definition of a unique null vector and of a single negative vector for each vector in it.

In many vector spaces the concept of *length* (specifying the 'magnitude') can be introduced, usually via an intermediate step. A vector space is called *Euclidean* if one can define for it a *scalar product* between two vectors, a number $\mathbf{ab}$ satisfying

$$\mathbf{aa} \geqslant 0 \,, \ \mathbf{ab} = \mathbf{ba} \,, \ (\mathbf{a}+\mathbf{a}')\mathbf{b} = \mathbf{ab}+\mathbf{a}'\mathbf{b} \,, \ \mathbf{a}(\mathbf{b}+\mathbf{b}') = \mathbf{ab}+\mathbf{ab}' \ \text{and} \ (c\mathbf{a})\mathbf{b} = \mathbf{a}(c\mathbf{b}) = c(\mathbf{ab}) . \tag{13}$$

In Cartesian coordinate notation, the standard scalar product is given by $\mathbf{ab} = a_x b_x + a_y b_y + a_z b_z$. Whenever it vanishes the two vectors are *orthogonal*. The *length* or *norm* of a vector can then be defined as the square root of the scalar product of a vector with itself: $a = \sqrt{\mathbf{aa}}$.

The scalar product is also useful to specify directions. Indeed, the scalar product between two vectors Challenge 101 n encodes the angle between them. Can you deduce this important relation?

** If I know an object I also know all its possible occurrences in states of affairs.

Ref. 43 *** Matter is a word derived from the Latin 'materia', which originally meant 'wood' and was derived via intermediate steps from 'mater', meaning 'mother'.

Table 10 Some measured acceleration values

Observation	Acceleration
Acceleration at Equator due to Earth's rotation	$0.34\,\mathrm{mm/s^2}$
Centrifugal acceleration due to the Earth's rotation	$33\,\mathrm{mm/s^2}$
Electron acceleration in household wire due to alternating current	$50\,\mathrm{mm/s^2}$
Gravitational acceleration on the Moon	$1.6\,\mathrm{m/s^2}$
Gravitational acceleration on the Earth's surface, depending on location	$9.8 \pm 0.1\,\mathrm{m/s^2}$
Standard gravitational acceleration	$9.806\,65\,\mathrm{m/s^2}$
Highest acceleration for a car or motor bike with engine-driven wheels	$15\,\mathrm{m/s^2}$
Gravitational acceleration on Jupiter's surface	$240\,\mathrm{m/s^2}$
Acceleration of cheetah	$32\,\mathrm{m/s^2}$
Acceleration that triggers air bags in cars	$360\,\mathrm{m/s^2}$
Fastest leg-powered acceleration (by the froghopper, *Philaenus spumarius*, an insect)	$4\,\mathrm{km/s^2}$
Tennis ball against wall	$0.1\,\mathrm{Mm/s^2}$
Bullet acceleration in rifle	$5\,\mathrm{Mm/s^2}$
Fastest centrifuges	$0.1\,\mathrm{Gm/s^2}$
Acceleration of protons in large accelerator	$90\,\mathrm{Tm/s^2}$
Acceleration of protons inside nucleus	$10^{31}\,\mathrm{m/s^2}$
Highest possible acceleration in nature	$10^{52}\,\mathrm{m/s^2}$

have a certain form or *shape*, defined as the precise way in which this impenetrability is distributed in space.

In order to describe motion as accurately as possible, it is convenient to start with those bodies that are as simple as possible. In general, the smaller a body, the simpler it is. A body that is so small that its parts no longer need to be taken into account is called a *particle*. (The older term *corpuscle* has fallen out of fashion.) Particles are thus idealized little stones. The extreme case, a particle whose size is *negligible* compared to the dimensions of its motion, so that its position is described completely by a *single* triplet of coordinates, is called a *point particle* or a *point mass*. In equation (5), the stone was assumed to be such a point particle.

Do point-like objects, i.e. objects smaller than anything one can measure, exist in daily life? Yes and no. The most notable examples are the stars. At present, angular sizes as small as 2 µrad can be measured, a limit given by the fluctuations of the air in the atmosphere. In space, such as for the Hubble telescope orbiting the Earth, the angular limit is due to the diameter of the telescope and is of the order of 10 nrad. Practically all stars seen from Earth are smaller than that, and are thus effectively 'point-like', even when seen with the most powerful telescopes.

As an exception to the general rule, the size of a few large and nearby stars, of red giant type, can be measured with special instruments.* Betelgeuse, the higher of the two

* The website http://www.astro.uiuc.edu/~kaler/sow/sowlist.html gives an introduction to the different

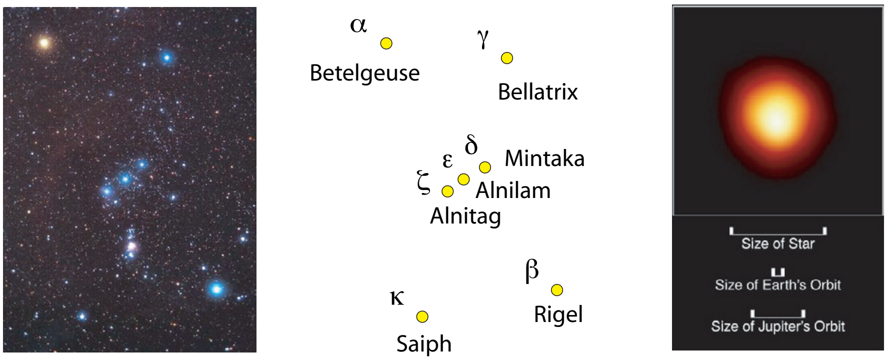


Figure 26 Orion (in natural colours) and Betelgeuse

shoulders of Orion shown in Figure 26, Mira in Cetus, Antares in Scorpio, Aldebaran in Taurus and Sirius in Canis Major are examples of stars whose size has been measured; they are all only a few light years from Earth. Of course, like the Sun, all other stars have a finite size, but one cannot prove this by measuring dimensions in photographs. (True?)

Ref. 44

Challenge 103 n

The difference between 'point-like' and finite size sources can be seen with the naked eye: at night, stars twinkle, but planets do not. (Check it!) This effect is due to the turbulence of air. Turbulence has an effect on the almost point-like stars because it deflects light rays by small amounts. On the other hand, air turbulence is too weak to lead to twinkling of sources of larger angular size, such as planets or artificial satellites,* because the deflection is averaged out in this case.

Challenge 104 e

An object is *point-like for the naked eye* if its angular size is smaller than about $2' = 0.6$ mrad. Can you estimate the size of a 'point-like' dust particle? By the way, an object is *invisible* to the naked eye if it is point-like *and* if its luminosity, i.e. the intensity of the light from the object reaching the eye, is below some critical value. Can you estimate whether there are any man-made objects visible from the Moon, or from the space shuttle?

Challenge 105 n

Challenge 106 n

The above definition of 'point-like' in everyday life is obviously misleading. Do proper, real point particles exist? In fact, is it possible at all to show that a particle has vanishing size? This question will be central in the last two parts of our walk. In the same way, we need to ask and check whether points in space do exist. Our walk will lead us to the astonishing result that all the answers to these questions are negative. Can you imagine how this could be proven? Do not be disappointed if you find this issue difficult; many brilliant minds have had the same problem.

Challenge 107 n

However, many particles, such as electrons, quarks or photons are point-like for all

types of stars. The http://www.astro.wisc.edu/~dolan/constellations/constellations.html web site provides detailed and interesting information about constellations.

For an overview of the planets, see the beautiful book by K.R. Lang & C.A. Whitney, *Vagabonds de l'espace – Exploration et découverte dans le système solaire*, Springer Verlag, 1993. The most beautiful pictures of the stars can be found in D. Malin, *A View of the Universe*, Sky Publishing and Cambridge University Press, 1993.

* A *satellite* is an object circling a planet, like the Moon; an *artificial satellite* is a system put into orbit by humans, like the Sputnik.

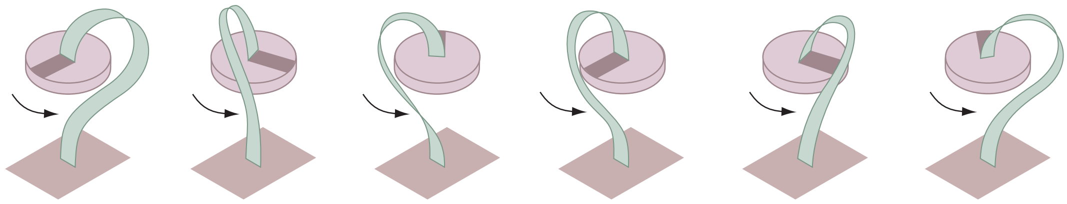


Figure 27 How an object can rotate continuously without tangling up the connection to a second one

practical purposes. Once one knows how to describe the motion of point particles, one can also describe the motion of extended bodies, rigid or deformable, by assuming that they are made of parts. This is the same approach as describing the motion of an animal as a whole by combining the motion of its various body parts. The simplest description, the *continuum approximation*, describes extended bodies as an infinite collection of point particles. It allows us to understand and to predict the motion of milk and honey, the motion of the air in hurricanes and of perfume in rooms. The motion of fire and all other gaseous bodies, the bending of bamboo in the wind, the shape changes of chewing gum, and the growth of plants and animals can also be described in this way. Ref. 45

Page 656 A more precise description than the continuum approximation is given below. Nevertheless, all observations so far have confirmed that the motion of large bodies can be described to high precision as the result of the motion of their parts. This approach will guide us through the first two parts of our mountain ascent. Only in the third part will we discover that, at a fundamental scale, this decomposition is impossible.

Legs and wheels

The parts of a body determine its shape. Shape is an important aspect of bodies: among others, it tells us how to count them. In particular, living beings are always made of a single body. This is not an empty statement: from this fact we can deduce that animals cannot have wheels or propellers, but only legs, fins, or wings. Why?

Living beings have only one surface; simply put, they have only one piece of skin. Mathematically speaking, animals are *connected*. This is often assumed to be obvious, and it is often mentioned that the blood supply, the nerves and the lymphatic connections to a rotating part would get tangled up. However, this argument is not correct, as Figure 27 shows. Can you find an example for this kind of motion in your own body? Are you able to see how many cables may be attached to the rotating body of the figure without hindering the rotation? In any case, the experiment shows that rotation is possible without tangling up connections. (Appendix D; Ref. 46; Challenge 108 n; Challenge 109 n)

Despite the possibility of rotating parts in animals, they still cannot be used to make a practical wheel or propeller. Can you see why? Evolution had no choice than to avoid animals with parts rotating around axes. Of course, this limitation does not rule out that living bodies move by rotation as a whole: the tumbleweed, seeds from various trees, some insects, certain other animals, children and dancers occasionally move by rolling or rotating as a whole. (Challenge 110 n; Ref. 47)

Single bodies, and thus all living beings, can only move through *deformation* of their

Figure still to be added

Figure 28 Legs and 'wheels' in living beings

shape: therefore they are limited to walking, running, crawling or flapping wings or fins, as shown in Figure 28. In contrast, systems of several bodies, such as bicycles, pedal boats or other machines, can move without any change of shape of their components, thus enabling the use of wheels, propellers or other rotating devices.*

In summary, whenever we observe a construction in which some part is turning continuously (and without the 'wiring' of the figure) we know immediately that it is an artefact: it is a machine, not a living being (but built by one). However, like so many statements about living creatures, this one also has exceptions. The distinction between one and two bodies is poorly defined if the whole system is made of only a few molecules. This happens most clearly inside bacteria. Organisms such as *Escherichia coli*, the well-known bacterium found in the human gut, or bacteria from the *Salmonella* family, all swim using flagella. *Flagella* are thin filaments, similar to tiny hairs sticking out of the cell membrane. In the 1970s it was shown that each flagellum, made of one or a few long molecules with a
Page 899 diameter of a few tens of nanometres, does in fact turn about its axis. A bacterium is able to turn its flagella in both clockwise and anticlockwise directions, can achieve more than
Ref. 48 1000 turns per second, and can turn all its flagella in perfect synchronization. Therefore wheels actually do exist in living beings, albeit only tiny ones. But let us now continue with our study of simple objects.

Objects and images

Walking through a forest we observe two rather different types of motion: the breeze moves the leaves, and at the same time their shadows move on the ground. Shadows
Ref. 49 are a simple type of image. Both objects and images are able to move. Running tigers, falling snowflakes, and material ejected by volcanoes are examples of motion, since they all change position over time. For the same reason, the shadow following our body, the beam of light circling the tower of a lighthouse on a misty night, and the rainbow that constantly keeps the same apparent distance from the hiker are examples of motion.

Everybody who has ever seen an animated cartoon in the cinema knows that images can move in more surprising ways than objects. Images can change their size, shape and

* Despite the disadvantage of not being able to use rotating parts and of being restricted to one piece only, nature's moving constructions, usually called animals, often outperform human built machines. As an example, compare the size of the smallest flying systems built by evolution with those built by humans. (See, e.g. http://pixelito.reference.be) The discrepancy has two reasons. First of all, nature's systems have integrated repair and maintenance systems. Second, nature can build large structures inside containers with small openings. In fact, nature is very good at building sailing ships inside glass bottles. The human body is
Challenge 111 n full of such examples; can you name a few?

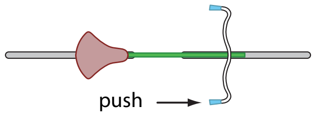


Figure 29 In which direction does the bicycle turn?

even colour, a feat only few objects are able to perform.* Images can appear and disappear without trace, multiply, interpenetrate, go backwards in time and defy gravity or any other force. Images, even ordinary shadows, can move faster than light. Images can float in space and keep the same distance to approaching objects. Objects cannot do almost
Ref. 51 anything of this. In general, the 'laws of cartoon physics' are rather different from those in nature. In fact, the motion of images does not seem to follow any rules at all, in contrast to the motion of objects. On the other hand, both objects and images differ from their environment in that they have *boundaries* defining their size and shape. We feel the need for precise criteria allowing the two cases to be distinguished.

The clearest distinction between images and objects is performed with the same method that children or animals use when they stand in front of a mirror for the first time: they try to *touch* what they see. Indeed, if we are able to touch what we see – or more precisely, if we are able to move it – we call it an *object*, otherwise an *image*.* Im-
Page 735 ages cannot be touched, but objects can. Images cannot hit each other, but objects can. And as everybody knows, touching something means to feel that it resists movement. Certain bodies, such as butterflies, pose little resistance and are moved with ease, others, such as ships, resist more, and are moved with more difficulty. This resistance to motion – more precisely, to change of motion – is called *inertia*, and the difficulty with which a body can be moved is called its *(inertial) mass*. Images have neither inertia nor mass.

Summing up, for the description of motion we must distinguish bodies, which can be touched and are impenetrable, from images, which cannot and are not. Everything visible
Challenge 113 n is either an object or an image; there is no third possibility. (Do you agree?) If the object is so far away that it cannot be touched, such as a star or a comet, it can be difficult to decide whether one is dealing with an image or an object; we will encounter this difficulty
Challenge 114 n repeatedly. For example, how would you show that comets are objects and not images?

In the same way that objects are made of *matter*, images are made of *radiation*. Images
Ref. 52 are the domain of shadow theatre, cinema, television, computer graphics, belief systems and drug experts. Photographs, motion pictures, ghosts, angels, dreams and many hallu-

* Excluding very slow changes such as the change of colour of leaves in the fall, in nature only certain crystals, the octopus, the chameleon and a few other animals achieve this. Of man-made objects, television,
Challenge 112 n computer displays, heated objects and certain lasers can do it. Do you know more examples? An excellent source of information on the topic of colour is the book by K. Nassau, *The Physics and Chemistry of Colour – the fifteen causes of colour*, J. Wiley & Sons, 1983. In the popular science domain, the most beautiful book is the classic work by the Flemish astronomer Marcel G.J. Minnaert, *Light and Colour in the Outdoors*, Springer, 1993, an updated version based on his wonderful book series, *De natuurkunde van 't vrije veld*,
Ref. 50 Thieme & Cie, Zutphen. Reading it is a must for all natural scientists. On the web, there is also the – much simpler – http://webexhibits.org/causesofcolour website.

* One could propose including the requirement that objects may be rotated; however, this requirement gives difficulties in the case of atoms, as explained on page 701, and with elementary particles, so that rotation is not made a separate requirement.

cinations are images (sometimes coupled with brain malfunction). To understand images, we need to study radiation (plus the eye and the brain). However, due to the importance of objects – after all we are objects ourselves – we study the latter first.

Motion and contact

> Democritus affirms that there is only one type of movement: That resulting from collision.
>
> Aetius, *Opinions*.

Ref. 53

When a child rides a monocycle, she or he makes use of a general rule in our world: one body acting on another puts it in motion. Indeed, in about six hours, anybody can learn to ride and enjoy it. As in all of life's pleasures, such as toys, animals, women, machines, children, men, the sea, wind, cinema, juggling, rambling and loving, something pushes something else. Thus our first challenge is to describe this transfer of motion in more precise terms.

Contact is not the only way to put something into motion; a counterexample is an apple falling from a tree or one magnet pulling another. Non-contact influences are more fascinating: nothing is hidden, but nevertheless something mysterious happens. Contact motion seems easier to grasp, and that is why one usually starts with it. However, despite this choice, non-contact forces are not easily avoided. Taking this choice one makes a similar experience that cyclists make. (See Figure 29.) When riding a bicycle at sustained speed and trying to turn left by pushing the right side of the steering bar, one takes a *right* turn.* In other words, despite our choice the rest of our walk will rapidly force us to study non-contact interactions as well.

What is mass?

> Δός μοι ποῦ στω καὶ κινῶ τὴν γῆν.
> Da ubi consistam, et terram movebo.**
>
> Archimedes

When we push something we do not know, such as when we kick an object on the street, we automatically pay attention to the same aspect that children explore when they stand before a mirror for the first time, or when they see a red laser spot for the first time. We check whether the unknown entity can be pushed and pay attention to how the unknown object moves under our influence. The high precision version of the experiment is shown in Figure 30. Repeating the experiment with various pairs of objects, we find – like in everyday life – that a fixed quantity m_i can be ascribed to every object i. The more it is difficult to move an object, the higher the quantity; it is determined by the relation

$$\frac{m_2}{m_1} = -\frac{\Delta v_1}{\Delta v_2} \tag{14}$$

* This surprising effect obviously works only above a certain minimal speed. Can you determine what this speed is? Be careful! Too strong a push will make you fall. Challenge 115 n

** 'Give me a place to stand, and I'll move the Earth.' Archimedes (*c.* 283–212), Greek scientist and engineer, is cited with this phrase by Pappus. Already Archimedes knew that the distinction used by lawyers between movable and immovable property makes no sense. Ref. 54

where Δv is the velocity change produced by the collision. The number m_i is called the *mass* of the object.

In order to get mass values common to everybody, the mass value for one particular, selected object has to be fixed in advance. This special object is called the *standard kilogram* and is kept with great care under vacuum in a glass container in Sèvres near Paris. It is touched only once every few years because otherwise dust, humidity, or scratches would change its mass. Through the standard kilogram the value of the mass of every other object in the world is determined.

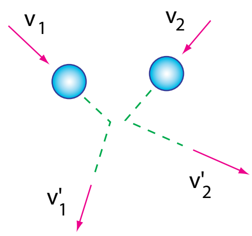


Figure 30 Collisions define mass

The mass thus measures the difficulty of getting something moving. High masses are harder to move than low masses. Obviously, only objects have mass; images don't. (By the way, the word 'mass' is derived, via Latin, from the Greek μαζα – bread – or the Hebrew 'mazza' – unleavened bread – quite a change in meaning.) Ref. 43

Experiments with everyday life objects also show that throughout any collision, the sum of all masses is conserved:

$$\sum_i m_i = \text{const} . \qquad (15)$$


Antoine Lavoisier

Conservation of mass was first stated by Antoine-Laurent Lavoisier.* Conservation of mass implies that the mass of a composite system is the sum of the mass of the components. In short, *Galilean mass is a measure for the quantity of matter.*

The definition of mass can also be given in another way. We can ascribe a number m_i to every object i such that for collisions free of outside interference the following sum is unchanged *throughout* the collision:

$$\sum_i m_i \mathbf{v}_i = \text{const} . \qquad (16)$$

The product of the velocity $\mathbf{v}_i$ and the mass m_i is called the *momentum* of the body. The sum, or *total momentum* of the system, is the same before and after the collision; it is a *conserved* quantity. *Momentum conservation defines mass.* The two conservation principles (15) and (16) were first stated in this way by the important Dutch physicist Christiaan Huygens.**

* Antoine-Laurent Lavoisier (1743–1794), French chemist and a genius. Lavoisier was the first to understand that combustion is a reaction with oxygen; he discovered the components of water and introduced mass measurements into chemistry. When he was (unjustly) sentenced to the guillotine during the French revolution, he decided to use the experience for a scientific experiment; he decided to blink his eyes as frequently as possible after his head was cut off, in order to show others how long it takes to lose consciousness. Lavoisier managed to blink eleven times.

** Christiaan Huygens (b. 1629 's Gravenhage, d. 1695 Hofwyck) was one of the main physicists and mathematicians of his time. Huygens clarified the concepts of mechanics; he also was one of the first to show that

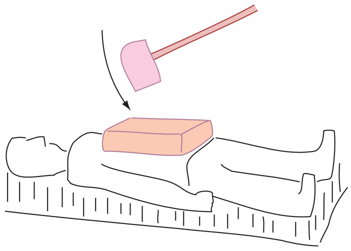
Christiaan Huygens

Momentum conservation implies that when a moving sphere hits a resting one of the same mass, a simple rule determines the angle between the directions the two spheres take after the collision. Can you find this rule? It is particularly useful when playing billiards.

Challenge 116 n

Another consequence is shown in Figure 31. A man lying on a bed of nails with two large blocks of concrete on his stomach. Another man is hitting the concrete with a heavy sledgehammer. As the impact is mostly absorbed by the concrete, there is no pain and no danger – unless the concrete is missed. Why?

Challenge 117 n



preliminary figure

Figure 31 Is this dangerous?

The above definition of mass has been generalized by the physicist and philosopher Ernst Mach* in such a way that it is valid even if the two objects interact without contact, as long as they do so along the line connecting their positions. The mass ratio between two bodies is defined as negative inverse acceleration ratio, thus as

$$\frac{m_2}{m_1} = -\frac{a_1}{a_2} , \tag{17}$$

where a is the acceleration of each body during the interaction. This definition has been studied in much detail in the physics community, mainly in the nineteenth century. A few points sum up the results:

- The definition of mass *implies* the conservation of momentum $\sum mv$. Momentum conservation is *not* a separate principle. Conservation of momentum cannot be checked experimentally, because mass is defined in such a way that it holds.
- The definition of mass *implies* the equality of the products $m_1 a_1$ and $-m_2 a_2$. Such products are called *forces*. The equality of acting and reacting forces is not a separate principle; mass is defined in such a way that it holds.

light is a wave. He wrote influential books on probability theory, clock mechanisms, optics and astronomy. Among other achievements, Huygens showed that the Orion Nebula consists of stars, discovered Titan, the moon of Saturn, and showed that the rings of Saturn consist of rock. (This is in contrast to Saturn itself, whose density is lower than that of water.)

* Ernst Mach (1838 Chrlice–1916 Vaterstetten), Austrian physicist and philosopher. The *mach* unit for aeroplane speed as a multiple of the speed of sound in air (about 0.3 km/s) is named after him. He developed the so-called Mach–Zehnder interferometer; he also studied the basis of mechanics. His thoughts about mass and inertia influenced the development of general relativity, and led to Mach's principle, which will appear later on. He was also proud to be the last scientist denying – humorously, and against all evidence – the existence of atoms.

Table 11 Properties of Galilean mass

Masses	Physical property	Mathematical name	Definition
Can be distinguished	distinguishability	element of set	Page 599
Can be ordered	sequence	order	Page 1099
Can be compared	measurability	metricity	Page 1108
Can change gradually	continuity	completeness	Page 1116
Can be added	quantity of matter	additivity	Page 66
Beat any limit	infinity	unboundedness, openness	Page 600
Do not change	conservation	invariance	$m = \text{const}$
Do not disappear	impenetrability	positivity	$m \geqslant 0$

▪ The definition of mass is *independent* of whether contact is involved or not, and whether the origin of the accelerations is due to electricity, gravitation, or other interactions.* Since the interaction does not enter the definition of mass, mass values defined with the help of the electric, nuclear or gravitational interaction all agree, as long as momentum is conserved. All known interactions conserve momentum. For some unfortunate historical reasons, the mass value measured with the electric or nuclear interactions is called the 'inertial' mass and the mass measured using gravity is called the 'gravitational' mass. As it turns out, this artificial distinction has no real meaning; this becomes especially clear when one takes an observation point that is far away from all the bodies concerned.

▪ The definition of mass is valid only for observers at rest or in inertial motion. More about this issue later on.

By measuring the masses of bodies around us, as given in Table 12, we can explore the science and art of experiments. We also discover the main properties of mass. It is *additive* in everyday life, as the mass of two bodies combined is equal to the sum of the two separate masses. Furthermore, mass is *continuous*; it can seemingly take any positive value. Finally, mass is *conserved*; the mass of a system, defined as the sum of the mass of all constituents, does not change over time if the system is kept isolated from the rest of the world. Mass is not only conserved in collisions but also during melting, evaporation, digestion and all other processes.

Later we will find that in the case of mass all these properties, summarized in Table 11, are only approximate. Precise experiments show that none of them are correct.* For the moment we continue with the present, Galilean concept of mass, as we have not yet a better one at our disposal.

* As mentioned above, only *central* forces obey the relation (17) used to define mass. Central forces act between the centre of mass of bodies. We give a precise definition later on. Page 84 But since all fundamental forces are central, this is not a restriction. There seems to be one notable exception: magnetism. Is the definition of mass valid in this case? Challenge 118 n

* In particular, in order to define mass we must be able to *distinguish* bodies. This seems a trivial requirement, but we discover that this is not always possible in nature.

Table 12 Some measured mass values

Observation	Mass
Mass increase due to absorption of one green photon	$3.7 \cdot 10^{-36}$ kg
Lightest known object: electron	$9.109\,381\,88(72) \cdot 10^{-31}$ kg
Atom of argon	$39.962\,383\,123(3)$ u = $66.359\,1(1)$ yg
Lightest object ever weighed (a gold particle)	0.39 ag
Human at early age (fertilized egg)	10^{-8} g
Water adsorbed onto a kilogram metal weight	10^{-5} g
Planck mass	$2.2 \cdot 10^{-5}$ g
Fingerprint	10^{-4} g
Typical ant	10^{-4} g
Water droplet	1 mg
Honey bee	0.1 g
Heaviest living things, such as the fungus *Armillaria ostoyae* or a large Sequoia *Sequoiadendron giganteum*	10^{6} kg
Largest ocean-going ship	$400 \cdot 10^{6}$ kg
Largest object moved by man (Troll gas rig)	$687.5 \cdot 10^{6}$ kg
Large antarctic iceberg	10^{15} kg
Water on Earth	10^{21} kg
Solar mass	$2.0 \cdot 10^{30}$ kg
Our galaxy	10^{41} kg
Total mass visible in the universe	10^{54} kg

In a famous experiment in the sixteenth century, for several weeks Santorio Santorio (Sanctorius) (1561–1636), friend of Galileo, lived with all his food and drink supply, and also his toilet, on a large balance. He wanted to test mass conservation. How did the measured weight change with time? Challenge 119 n

The definition of mass through momentum conservation implies that during the fall of an object, the Earth is accelerated upwards by a tiny amount. If one could measure this tiny amount, one could determine the mass of the Earth. Unfortunately, this measurement is impossible. Can you find a better way to determine the mass of the Earth? Challenge 120 n

Summarizing Table 11, the mass of a body is thus most precisely described by a *positive* real number, often abbreviated m or M. This is a direct consequence of the impenetrability of matter. Indeed, a *negative* (inertial) mass would mean that such a body would move in the opposite direction of any applied force or acceleration. Such a body could not be kept in a box; it would break through any wall trying to stop it. Strangely enough, negative mass bodies would still fall downwards in the field of a large positive mass (though more slowly than an equivalent positive mass). Are you able to confirm this? Challenge 121 e However, a small positive mass object would float away from a large negative mass body, as you can easily deduce by comparing the various accelerations involved. A positive and a negative mass of the same value would stay at constant distance and spontaneously accelerate away along the line connecting the two masses. Challenge 122 e Note that both energy and momentum

are conserved in all these situations.* Negative-mass bodies have never been observed.
Page 287, page 708 Antimatter, which will be discussed later, also has positive mass.

Is motion eternal?

> Every body continues in the state of rest or of uniform motion in a straight line except in so far as it doesn't.
>
> Arthur Eddington*

The product $\mathbf{p} = m\mathbf{v}$ of mass and velocity is called the *momentum* of a particle; it describes the tendency of an object to keep moving during collisions. The bigger it is, the harder it is to stop the object. Like velocity, momentum has a direction and a magnitude: it is a vector. In French, momentum is called 'quantity of motion', a more appropriate term. In the old days, the term 'motion' was used instead of 'momentum', for example by Newton. Relation (16), the conservation of momentum, therefore expresses the conservation of motion during interactions.

Momentum and energy are *extensive quantities*. That means that it can be said of both that they *flow* from one body to the other, and that they can be *accumulated* in bodies, in the same way that water flows and can be accumulated in containers. Imagining momentum as something that can be *exchanged* between bodies in collisions is always useful when thinking about the description of moving objects.

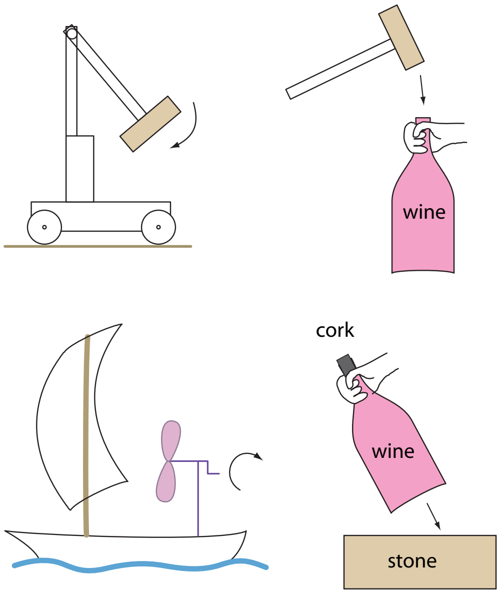


Figure 32 What happens?

Momentum is conserved. That explains the limitations you might experience when being on a perfectly frictionless surface, such as ice or a polished, oil covered marble: you cannot propel yourself forward by patting your own back. (Have you ever tried to put a cat on such a marble surface? It is not even able to stand on its four legs. Neither are humans. Can you imagine why?)
Challenge 125 n

The conservation of momentum and mass also means that teleportation ('beam me up') is impossible in nature. Can you explain this to a non-physicist?
Challenge 126 n

* For more curiosities, see R.H. Price, Negative mass can be positively amusing, *American Journal of Physics* **61**, pp. 216–217, 1993. Negative mass particles in a box would heat up a box made of positive mass while traversing its walls, and accelerating, i.e. losing energy, at the same time. They would allow building of a *perpetuum mobile* of the second kind, i.e. a device circumventing the second principle of thermodynamics. Moreover, such a system would have no thermodynamic equilibrium, because its energy could decrease forever. The more one thinks about negative mass, the more one finds strange properties contradicting observations. By the way, what is the range of possible mass values for tachyons?
Page 78
Challenge 123 e
Challenge 124 n

* Arthur Eddington (1882–1944), British astrophysicist.

Momentum conservation implies that momentum can be imagined like an invisible fluid. In an interaction, the invisible fluid is transferred from one object to another. However, the sum is always constant.

Momentum conservation implies that motion never stops; it is only *exchanged*. On the other hand, motion often disappears in our environment, as in the case of a stone dropped to the ground, or of a ball left rolling on grass. Moreover, in daily life we often observe creation of motion, such as every time we open a hand. How do these examples fit with the conservation of momentum?

It turns out that the answer lies in the microscopic aspects of these systems. A muscle only *transforms* one type of motion, namely that of the electrons in certain chemical compounds* into another, the motion of the fingers. The working of muscles is similar to that of a car engine transforming the motion of electrons in the fuel into motion of the wheels. Both systems need fuel and get warm in the process.

We must also study the microscopic behaviour when a ball rolls on grass until it stops. The disappearance of motion is called *friction*. Studying the situation carefully, one finds that the grass and the ball heat up a little during this process. During friction, visible motion is transformed into heat. Later, when we discover the structure of matter, it will become clear that heat is the disorganized motion of the microscopic constituents of every material. When these constituents all move in the same direction, the object as a whole moves; when they oscillate randomly, the object is at rest, but is warm. Heat is a form of motion. Friction thus only seems to be disappearance of motion; in fact it is a transformation of ordered into unordered motion.

Despite momentum conservation, *macroscopic* perpetual motion does not exist, since friction cannot be eliminated completely.** Motion is eternal only at the microscopic scale. In other words, the disappearance and also the spontaneous appearance of motion in everyday life is an illusion due to the limitations of our senses. For example, the motion proper to every living being exists before its birth, and stays after its death. The same happens with its energy. This result is probably the closest one can get to the idea of everlasting life from evidence collected by observation. It is perhaps less than a coincidence that energy used to be called *vis viva*, or 'living force', by Leibniz and many others.

Since motion is conserved, it has no origin. Therefore, at this stage of our walk we can-

Ref. 55 * Usually adenosine triphosphate (ATP), the fuel of most processes in animals.

** Some funny examples of past attempts to built a *perpetual motion machine* are described in STANISLAV MICHEL, *Perpetuum mobile*, VDI Verlag, 1976. Interestingly, the idea of eternal motion came to Europe from India, via the Islamic world, around the year 1200, and became popular as it opposed the then standard view that all motion on Earth disappears over time. See also the http://www.geocities.com/mercutio78_99/pmm.html and the http://www.lhup.edu/~dsimanek/museum/unwork.htm websites. The conceptual mistake made by all eccentrics and used by all crooks is always the same: the hope of overcoming friction. (In fact, this applied only to the perpetual motion machines of the second kind; those of the first kind – which are even more in contrast with observation – even try to generate energy from nothing.)

If the machine is well constructed, i.e. with little friction, it can take the little energy it needs for the sustenance of its motion from very subtle environmental effects. For example, in the Victoria and Albert Ref. 56 Museum in London one can admire a beautiful clock powered by the variations of air pressure over time.

Small friction means that motion takes a long time to stop. One immediately thinks of the motion of the Challenge 127 n planets. In fact, there *is* friction between the Earth and the Sun. (Can you guess one of the mechanisms?) But the value is so small that the Earth has already circled around the Sun for thousands of millions of years, and will do so for quite some time more.

not at all answer the fundamental questions: Why does motion exist? What is its origin? The end of our adventure is nowhere near.

More on conservation – energy

When collisions are studied in detail, a second conserved quantity turns up. Experiments show that in the case of perfect, or elastic collisions – collisions without friction – the following quantity, called the *kinetic energy* T of the system, is also conserved:

$$T = \sum_i \tfrac{1}{2} m_i \mathbf{v}_i^2 = \sum_i \tfrac{1}{2} m_i v_i^2 = \text{const} . \tag{18}$$

Kinetic energy thus depends on the mass and on the square of the speed v of a body. Kinetic energy is the ability a body has to induce change in bodies it hits. The full name 'kinetic energy' was introduced by Gustave-Gaspard Coriolis.* Coriolis also introduced the factor 1/2, in order that the relation $dT/dv = p$ would be obeyed. (Why?) Energy is a word taken from ancient Greek; originally it was used to describe character, and meant 'intellectual or moral vigour'. It was taken into physics by Thomas Young (1773–1829) in 1807 because its literal meaning is 'force within'. (The letters E, W, A and several others are also used to denote energy.) Another, equivalent definition of energy will become clear later on: energy is what can be transformed into heat.

Challenge 128 n

(Physical) *energy* is the measure of the ability to generate motion. A body has a lot of energy if it has the ability to move many other bodies. Energy is a number; it has no direction. The total momentum of two equal masses moving with opposite velocities is zero; their total energy increases with either velocity value. Energy thus also measures motion, but in a different way than momentum does. Energy measures motion in a more global way.

Do not be surprised if you do not grasp the difference between momentum and energy straight away: physicists took about two centuries to figure it out. For some time they even insisted on using the same word for both of them, and often they didn't know which situation required which concept. So you are allowed to take a few minutes to get used to it.

Both energy and momentum measure how systems change. Momentum tells how systems change over distance, energy measures how systems change over time. Momentum is needed to compare motion here and there. Energy is needed to compare motion now and later.

One way to express the difference between energy and momentum is to think about the following challenges. Is it more difficult to stop a running man with mass m and speed v, or one with mass $m/2$ and speed $2v$, or one with mass $m/2$ and speed $\sqrt{2}v$? You may want to ask a rugby-playing friend for confirmation.

Challenge 129 e

Another distinction is taught by athletics: the *real* long jump world record, almost 10 m, is still kept by an athlete who in the early twentieth century ran with two weights in his hands, and then threw the weights behind him in the moment he took off. Can you explain the feat?

Challenge 130 n

* Gustave-Gaspard Coriolis (b. 1792 Paris, d. 1843 Paris), French engineer and mathematician.

Table 13 Some measured energy values

Observation	Energy
Average kinetic energy of oxygen molecule in air	$6 \cdot 10^{-21}$ J
Green photon energy	$5.6 \cdot 10^{-20}$ J
X-ray photon energy	10^{-15} J
γ photon energy	10^{-12} J
Highest particle energy in accelerators	10^{-7} J
Comfortably walking human	20 J
Flying arrow	50 J
Right hook in boxing	50 J
Energy in torch battery	1 kJ
Flying rifle bullet	10 kJ
Apple digestion	0.2 MJ
Car on highway	1 MJ
Highest laser pulse energy	1.8 MJ
Lightning flash	up to 1 GJ
Planck energy	2.0 GJ
Small nuclear bomb (20 kton)	84 TJ
Earthquake of magnitude 7	2 PJ
Largest nuclear bomb (50 Mton)	210 PJ
Impact of meteorite with 2 km diameter	1 EJ
Yearly machine energy use	420 EJ
Rotation energy of Earth	$2 \cdot 10^{29}$ J
Supernova explosion	10^{44} J
Gamma ray burst	up to 10^{47} J
Energy content $E = mc^2$ of Sun's mass	$1.8 \cdot 10^{47}$ J
Energy content of Galaxy's central black hole	$4 \cdot 10^{53}$ J

When a car travelling at 100 m/s runs frontally into a parked car of the same make, which car receives the larger damage? What changes if the parked car has its brakes on? Challenge 131 n

To get a better feeling for energy, here is an additional approach. The world use of energy by human machines (coming from solar, geothermal, biomass, wind, nuclear, hydro, gas, oil, coal, or animal sources) in the year 2000 is about 420 EJ,* for a world population of about 6000 million people. Ref. 57 To see what this energy consumption means, we translate it into a personal power consumption; we get about 2.2 kW. The Watt W is the unit of power, and is simply defined as 1 W = 1 J/s, reflecting the definition of *(physical) power* as energy used per unit time. As a working person can produce mechanical work of about 100 W, the average human energy consumption corresponds to

Robert Mayer

Page 1060 * For the explanation of the abbreviation E, see Appendix B.

about 22 humans working 24 hours a day. In particular, if we look at the energy consumption in countries of the First World, the average inhabitant there has machines working for him equivalent to several hundred 'servants'. Can you point out some of these machines?

Challenge 132 n

Kinetic energy is thus not conserved in everyday life. For example, in non-elastic collisions, like that of a piece of chewing gum hitting a wall, kinetic energy is lost. *Friction* destroys kinetic energy, as it destroys momentum. At the same time, friction produces heat. It was one of the important conceptual discoveries of physics that *total* energy is conserved if one includes the discovery that heat is a form of energy. Friction is thus in fact a process transforming kinetic energy, i.e. the energy connected with the motion of a body, into heat. On a microscopic scale, energy is conserved.* Indeed, without energy conservation, the concept of time would not be definable. We will show this connection shortly.

Is velocity absolute? – The theory of everyday relativity

Why don't we feel all the motions of the Earth? The two parts of the answer were already given in 1632. First of all, as Galileo explained, we do not feel the accelerations of the Earth because the effects they produce are too tiny to be detected by our senses. Indeed, many of the mentioned accelerations do induce measurable effects in high-precision experiments, e.g. in atomic clocks.

Page 359

But the second point made by Galileo is equally important. We do not feel translational, unaccelerated motions because this is impossible *in principle*. We cannot feel that we move! Galileo discussed the issue by comparing the observations of two observers: one on ground and another on the most modern means of transportation of the time, a ship. Galileo asked whether a man on the ground and a man in a ship moving at constant speed experience (or 'feel') anything different. Einstein used observers in trains. Later it became fashionable to use travellers in rockets. (What will come next?) Galileo explained that only *relative* velocities between bodies produce effects, not the absolute values of the velocities. For the senses, there is no difference between constant, undisturbed motion, however rapid it may be, and rest. This is now called *Galileo's principle of relativity*. In everyday life we feel motion only if the means of transportation trembles (thus if it accelerates), or if we move against the air. Therefore Galileo concludes that two observers in straight and undisturbed motion against each other cannot say who is 'really' moving. Whatever their relative speed, neither of them 'feels' in motion.**

Challenge 133 e

* In fact, the conservation of energy was stated in its full generality in public only in 1842, by Julius Robert Mayer. He was a medical doctor by training, and the journal *Annalen der Physik* refused to publish his paper, as it supposedly contained 'fundamental errors.' What the editors called errors were in fact mostly – but not only – contradictions of their prejudices. Later on, Helmholtz, Kelvin, Joule and many others acknowledged Mayer's genius. However, the first to have stated energy conservation in its modern form was the French physicist Sadi Carnot (1796–1832) in 1820. To him the issue was so clear that he did not publish the result. In fact he went on and discovered the second 'law' of thermodynamics. Today, energy conservation, also called the first 'law' of thermodynamics, is one of the pillars of physics, as it is valid in all its domains.

** In 1632, in his *Dialogo*, Galileo writes: 'Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a wide vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all

Table 14 Some measured power values

Observation	Power
Power of flagellar motor in bacterium	0.1 pW
Incandescent light bulb light output	1 to 5 W
Incandescent light bulb electricity consumption	25 to 100 W
A human, during one work shift	100 W
One horse, for one shift	300 W
Eddy Merckx, the great bicycle athlete, during one hour	500 W
Official horse power	735 W
Large motor bike	100 kW
Electrical power station output	0.1 to 6 GW
World's electrical power production in 2000	450 GW
Power used by the geodynamo	200 to 500 GW
Input on Earth surface: Sun's irradiation of Earth Ref. 58	0.17 EW
Input on Earth surface: thermal energy from inside of Earth	32 TW
Input on Earth surface: power from tides (i.e. from Earth's rotation)	3 TW
Input on Earth surface: power generated by man from fossil fuels	8 to 11 TW
Lost from Earth surface: power stored by plants' photosynthesis	40 TW
World's record laser power	1 PW
Output of Earth surface: sunlight reflected into space	0.06 EW
Output of Earth surface: power radiated into space at 287 K	0.11 EW
Sun's output	384.6 YW
Maximum power in nature, $c^4/4G$	$9.1 \cdot 10^{51}$ W

sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to your friend, you need throw it no more strongly in one direction than another, the distances being equal: jumping with your feet together, you pass equal spaces in every direction. When you have observed all these things carefully (though there is no doubt that when the ship is standing still everything must happen in this way), have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that, you will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still. In jumping, you will pass on the floor the same spaces as before, nor will you make larger jumps toward the stern than toward the prow even though the ship is moving quite rapidly, despite the fact that during the time you are in the air the floor under you will be going in a direction opposite to your jump. In throwing something to your companion, you will need no more force to get it to him whether he is in the direction of the bow or the stern, with yourself situated opposite. The droplets will fall as before into the vessel beneath without dropping toward the stern, although while the drops are in the air the ship runs many spans. The fish in their water will swim toward the front of their bowl with no more effort than toward the back, and will go with equal ease to bait placed anywhere around the edges of the bowl. Finally the butterflies and flies will continue their flights indifferently toward every side, nor will it ever happen that they are concentrated toward the stern, as if tired out from keeping up with the course of the ship, from which they will have been separated during long intervals by keeping themselves in the air. And if smoke is made by burning some incense, it will be seen going up in the form of a little cloud, remaining still and moving no more toward one side than the other. The cause of all these correspondences of effects is the fact that the ship's motion is common to all the things

Rest is relative. Or more clearly: rest is an observer-dependent concept. This result of Galilean physics is so important that Poincaré introduced the expression 'theory of relativity' and Einstein repeated the principle explicitly when he published his famous theory of special relativity. However, these names are awkward. Galilean physics is also a theory of relativity! The relativity of rest is common to *all* of physics; it is an essential aspect of motion.

Undisturbed or uniform motion has no observable effect; only change of motion does. As a result, every physicist can deduce something simple about the following statement by Wittgenstein

> Daß die Sonne morgen aufgehen wird, ist eine Hypothese; und das heißt: wir wissen nicht, ob sie aufgehen wird.*

The statement is *wrong*. Can you explain why Wittgenstein erred here, despite his strong desire not to?

Challenge 134 n

Rotation

Rotation keeps us alive. Without the change of day and night, we would be either fried or frozen to death, depending on our location on our planet. A short summary of rotation is thus appropriate. We saw before that a body is described by its reluctance to move; similarly, a body also has a reluctance to turn. This quantity is called its *moment of inertia* and is often abbreviated Θ. The speed or rate of rotation is described by *angular velocity*, usually abbreviated ω. A few values found in nature are given in Table 15.

The observables that describe rotation are similar to those describing linear motion, as shown in Table 16. Like mass, the moment of inertia is defined in such a way that the sum of *angular momenta* L – the product of moment of inertia and angular velocity – is conserved in systems that do not interact with the outside world:

$$\sum_{\mathrm{i}} \Theta_{\mathrm{i}} \omega_{\mathrm{i}} = \sum_{\mathrm{i}} L_{\mathrm{i}} = \text{const} \tag{19}$$

In the same way that linear momentum conservation defines mass, angular momentum conservation defines the moment of inertia.

The moment of inertia can be related to the mass and shape of a body. If the body is imagined to consist of small parts or mass elements, the resulting expression is

$$\Theta = \sum_{\mathrm{n}} m_{\mathrm{n}} r_{\mathrm{n}}^2 , \tag{20}$$

where r_{n} is the distance from the mass element m_{n} to the axis of rotation. Can you confirm the expression? Therefore, the moment of inertia of a body depends on the chosen axis

Challenge 135 e

contained in it, and to the air also. That is why I said you should be below decks; for if this took place above in the open air, which would not follow the course of the ship, more or less noticeable differences would be seen in some of the effects noted.'

* 'It is an hypothesis that the Sun will rise tomorrow; and this means that we do not *know* whether it will rise.' This well-known statement is found in Ludwig Wittgenstein, *Tractatus*, 6.36311.

Table 15 Some measured rotation frequencies

Observation	Angular velocity $\omega = 2\pi/T$
Galactic rotation	$2\pi \cdot 0.14 \cdot 10^{-15}/\mathrm{s} = 2\pi\,/220 \cdot 10^6$ a
Average Sun rotation around its axis	$2\pi \cdot 3.8 \cdot 10^{-7}/\mathrm{s} = 2\pi\,/\,30$d
Typical lighthouse	$2\pi \cdot 0.08/\mathrm{s}$
Jumping ballet dancer	$2\pi \cdot 3/\mathrm{s}$
Ship's diesel engine	$2\pi \cdot 5/\mathrm{s}$
Helicopter motor	$2\pi \cdot 5.3/\mathrm{s}$
Washing machine	up to $2\pi \cdot 20/\mathrm{s}$
Bacterial flagella	$2\pi \cdot 100/\mathrm{s}$
Racing car engine	up to $2\pi \cdot 600/\mathrm{s}$
Fastest turbine built	$2\pi \cdot 10^3/\mathrm{s}$
Fastest pulsars (rotating stars)	up to $2\pi \cdot 10^3/\mathrm{s}$
Ultracentrifuge	$> 2\pi \cdot 2 \cdot 10^3/\mathrm{s}$
Proton rotation	$2\pi \cdot 10^{20}/\mathrm{s}$
Highest possible, Planck angular velocity	$2\pi \cdot 10^{35}/\mathrm{s}$

Table 16 Correspondence between linear and rotational motion

Quantity	Linear motion		Rotation	
State	time	t	time	t
	position	x	angle	φ
	momentum	$p = mv$	angular momentum	$L = \Theta\omega$
	energy	$mv^2/2$	energy	$\Theta\omega^2/2$
Motion	velocity	v	angular velocity	ω
	acceleration	a	angular acceleration	α
Reluctance to move	mass	m	moment of inertia	Θ
Motion change	force	ma	torque	$\Theta\alpha$

Challenge 136 n of rotation. Can you confirm this for a brick?

Obviously, the value of the moment of inertia also depends on the location of the axis used for its definition. For each axis direction, one distinguishes *intrinsic* moment of inertia, when the axis passes through the centre of mass of the body, from *extrinsic* moment of inertia, when it does not.* In the same way, one distinguishes intrinsic and extrinsic angular momenta. (By the way, the *centre of mass* of a body is that imaginary

* Extrinsic and intrinsic moment of inertia are related by

$$\Theta_{\mathrm{ext}} = \Theta_{\mathrm{int}} + md^2 \,, \tag{21}$$

where d is the distance between the centre of mass and the axis of extrinsic rotation. This relation is called
Challenge 137 n *Steiner's parallel axis theorem*. Are you able to deduce it?

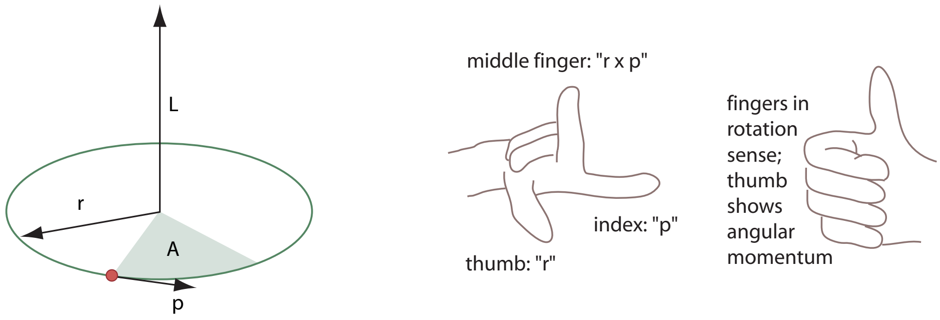


Figure 33 Angular momentum and the two versions of the right-hand rule

point which moves straight during vertical fall, even if the body is rotating. Can you find a way to determine its location for a specific body?)
Challenge 138 n

Every object that has an orientation also has an intrinsic angular momentum. (What about a sphere?) Therefore, point particles do not have intrinsic angular momenta – at least in first approximation. (This conclusion will change in quantum theory.) The *extrinsic* angular momentum **L** of a point particle is given by
Challenge 139 n

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} = \frac{2\mathbf{A}(T)m}{T} \quad \text{so that} \quad L = r\,p = \frac{2A(T)m}{T} \tag{22}$$

where **p** is the momentum of the particle, $\mathbf{A}(T)$ is the surface swept by the position vector **r** of the particle during time T.* The angular momentum thus points along the rotation axis, following the right hand rule, as shown in Figure 33.

We then define a corresponding *rotational energy* as

$$E_{\text{rot}} = \tfrac{1}{2}\Theta\,\omega^2 = \frac{L^2}{2\Theta}\ . \tag{24}$$

The expression is similar to the expression for the kinetic energy of a particle. Can you guess how much larger the rotational energy of the Earth is compared to the yearly electricity usage of humanity? In fact, if you can find a way to harness this energy, you will
Challenge 142 n

* For the curious, the result of the *cross product* or *vector product* $\mathbf{a}\times\mathbf{b}$ between two vectors **a** and **b** is defined as that vector that is orthogonal to both, whose orientation is given by the *right-hand rule*, and whose length is given by $ab \sin \sphericalangle(\mathbf{a},\mathbf{b})$, i.e. by the surface area of the parallelogram spanned by the two vectors. From the definition you can show that the vector product has the properties
Challenge 140 e

$$\begin{aligned}
&\mathbf{a}\times\mathbf{b} = -\mathbf{b}\times\mathbf{a} \quad , \quad \mathbf{a}\times(\mathbf{b}+\mathbf{c}) = \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \quad , \quad \lambda\mathbf{a}\times\mathbf{b} = \lambda(\mathbf{a}\times\mathbf{b}) = \mathbf{a}\times\lambda\mathbf{b} \quad , \quad \mathbf{a}\times\mathbf{a} = \mathbf{0} \quad , \\
&\mathbf{a}(\mathbf{b}\times\mathbf{c}) = \mathbf{b}(\mathbf{c}\times\mathbf{a}) = \mathbf{c}(\mathbf{a}\times\mathbf{b}) \quad , \quad \mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = \mathbf{b}(\mathbf{ac}) - \mathbf{c}(\mathbf{ab}) \quad , \\
&(\mathbf{a}\times\mathbf{b})(\mathbf{c}\times\mathbf{d}) = \mathbf{a}(\mathbf{b}\times(\mathbf{c}\times\mathbf{d})) = (\mathbf{ac})(\mathbf{bd}) - (\mathbf{bc})(\mathbf{ad}) \quad , \\
&(\mathbf{a}\times\mathbf{b})\times(\mathbf{c}\times\mathbf{d}) = \mathbf{c}((\mathbf{a}\times\mathbf{b})\mathbf{d}) - \mathbf{d}((\mathbf{a}\times\mathbf{b})\mathbf{c}) \quad , \quad \mathbf{a}\times(\mathbf{b}\times\mathbf{c}) + \mathbf{b}\times(\mathbf{c}\times\mathbf{a}) + \mathbf{c}\times(\mathbf{a}\times\mathbf{b}) = 0\ .
\end{aligned} \tag{23}$$

The vector product exists (almost) only in three-dimensional vector spaces. (See Appendix D) The cross product vanishes if and only if the vectors are parallel. The parallelepiped spanned by three vectors **a**, **b** and **c** has the volume $V = \mathbf{c}(\mathbf{a}\times\mathbf{b})$. The pyramid or tetrahedron formed by the three vectors has one sixth of that volume.
Page 1106
Challenge 141 e

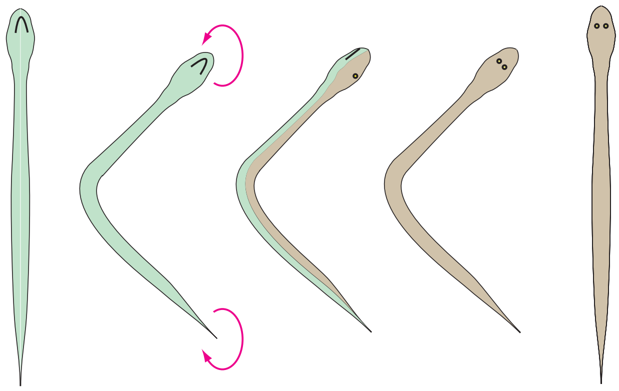


Figure 34 How a snake turns itself around its axis

become famous.

As in the case of linear motion, rotational energy and angular momentum are not always conserved in the macroscopic world, due to friction; but they are always conserved on the microscopic scale.

On a frictionless surface, as approximated by smooth ice or by a marble floor covered by a layer of oil, it is impossible to move forward. In order to move, we need to push *against* something. Is this also the case for rotation?

Surprisingly, it is possible to turn even *without* pushing against something. You can check this on a well-oiled rotating office chair: simply rotate an arm above the head. After each turn of the hand, the orientation of the chair has changed by a small amount. Indeed, conservation of angular momentum and of rotational energy do not prevent bodies from changing their orientation. Cats learn this in their youth. After they have learned the trick, if they are dropped legs up, they can turn themselves in such a way that they always land feet first. Snakes also know how to rotate themselves, as Figure 34 shows. During the Olympic Games one can watch board divers and gymnasts perform similar tricks. Rotation is thus different from translation in this aspect. (Why?)

Ref. 59

Challenge 143 d

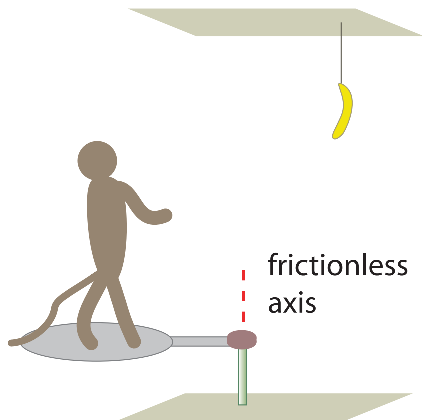


Figure 35 Can the ape reach the banana?

Angular momentum is conserved. This statement is valid for any axis, provided that friction plays no role. To make the point, Jean-Marc Lévy-Leblond poses the problem of Figure 35. Can the ape reach the banana without leaving the plate, assuming that the plate on which the ape rests can turn around the axis without friction?

Ref. 2

Challenge 144 n

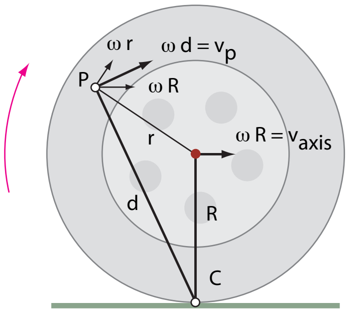


Figure 36 The velocities and unit vectors for a rolling wheel

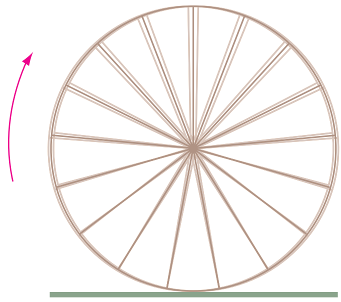


Figure 37 A simulated photograph of a rolling wheel with spokes

Rolling wheels

Rotation is an interesting phenomenon in many ways. A rolling wheel does *not* turn around its axis, but around its point of contact. Let us show this.

A wheel of radius R is *rolling* if the speed of the axis v_{axis} is related to the angular velocity ω by

$$\omega = \frac{v_{axis}}{R} . \tag{25}$$

For any point P on the wheel, with distance r from the axis, the velocity v_P is the sum of the motion of the axis and the motion around the axis. Figure 36 shows that v_P is orthogonal to d, the distance between the point P and the contact point of the wheel.

Challenge 145 e

The figure also shows that the length ratio between v_P and d is the same as between v_{axis} and R. As a result, we can write

$$\mathbf{v}_P = \omega \times \mathbf{d} \tag{26}$$

which shows that a rolling wheel does indeed rotate about its contact point with the ground.

Surprisingly, when a wheel rolls, some points on it move towards the wheel's axis, some stay at a fixed distance and others move away from it. Can you determine where these various points are located? Together, they lead to an interesting pattern when a rolling wheel with spokes, such as a bicycle wheel, is photographed.

Challenge 146 n

Ref. 60

Ref. 61

With these results you can tackle the following beautiful challenge. When a turning bicycle wheel is deposed on a slippery surface, it will slip for a while and then end up rolling. How does the final speed depend on the initial speed and on the friction?

Challenge 147 d

How do we walk?

> Golf is a good walk spoiled.
>
> Mark Twain

Why do we move our arms when walking or running? To conserve energy. In fact, when a body movement is performed with as little energy as possible, it is natural and graceful.

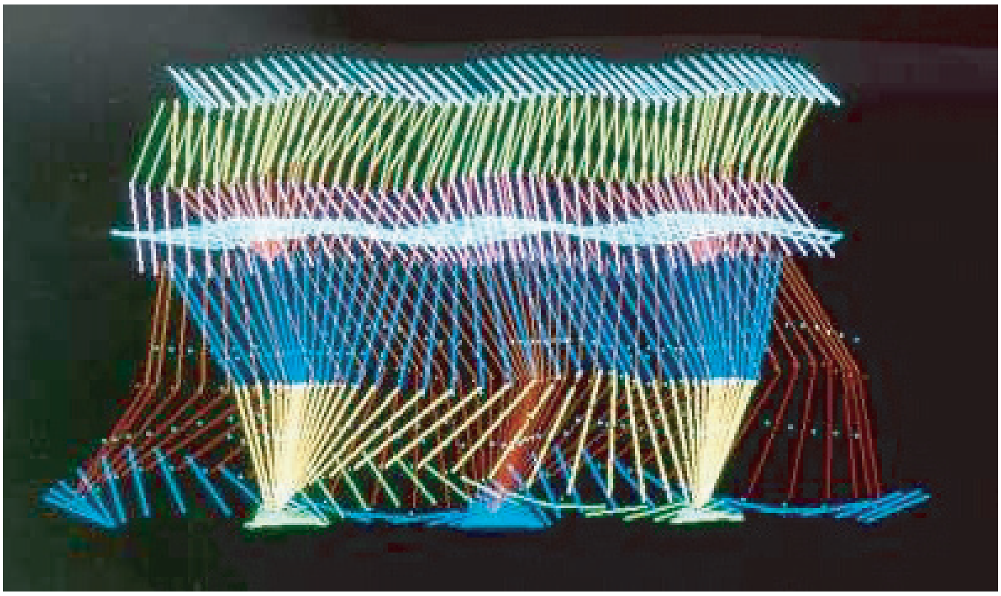


Figure 38 The measured motion of a walking human

(This can indeed be taken as the actual definition of grace. The connection is common knowledge in the world of dance; it is also a central aspect of the methods used by actors to learn how to move their bodies as beautifully as possible.) Ref. 14

To convince yourself about the energy savings, try walking or running with your arms fixed or moving in the opposite direction than usual: the effort required is considerably higher. In fact, when a leg is moved, it produces a torque around the body axis which has to be counterbalanced. The method using the least energy is the swinging of arms. Since the arms are lighter than the legs, they must move further from the axis of the body, to compensate for the momentum; evolution has therefore moved the attachment of the arms, the shoulders, farther away than those of the legs, the hips. Animals on two legs but without arms, such as penguins or pigeons, have more difficulty walking; they have to move their whole torso with every step.

Which muscles do most of the work when walking, the motion that experts call *gait*? In 1980, Serge Gracovetsky found that in human gait most power comes from the *spine* muscles, not from the legs. (Indeed, people without legs are also able to walk.) When you take a step, the lumbar muscles straighten the spine; this automatically makes it turn a bit to one side, so that the knee of the leg on that side automatically comes forward. When the foot is moved, the lumbar muscles can relax, and then straighten again for the next step. In fact, one can experience the increase in tension in the *back* muscles when walking without moving the arms, thus confirming where the human engine is located. Ref. 62 Challenge 148 e

Human legs differ from that of apes in a fundamental aspect: humans are able to run. In fact the whole human body has been optimized for running, an ability that no other primate has. The human body has shed most hair to achieve better cooling, has evolved the ability to run while keeping the head stable, has evolved the right length of arms for proper balance when running, and even has a special ligament in the back that works as

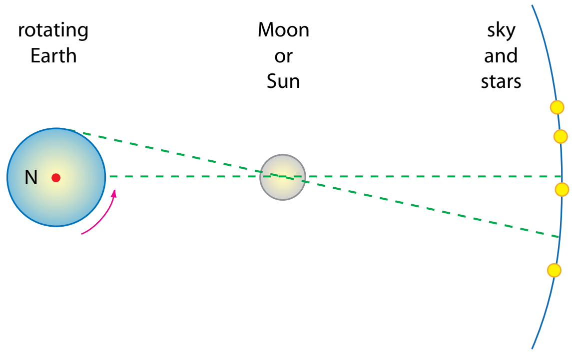


Figure 39 The parallaxis – not drawn to scale

a shock absorber while running. In other words, running is the most human of all forms of motion.

Is the Earth rotating?

Eppur si muove!

Anonymous*

The search for answers to this question gives a beautiful cross-section of the history of classical physics. Around the year 265 BCE, the Greek thinker Aristarchos of Samos maintained that the Earth rotates. He had measured the parallax of the Moon (today known to be up to 0.95°) and of the Sun (today known to be 8.8′).** The *parallax* is an interesting effect; it is the angle describing the difference between the directions of a body in the sky when seen by an observer on the surface of the Earth and when seen by a hypothetical observer at its centre. (See Figure 39.) Aristarchos noticed that the Moon and the Sun *wobble* across the sky, and this wobble has a period of 24 hours. He concluded that the Earth rotates.

Ref. 63

Measurements of the aberration of light also show the rotation of the Earth; it can be detected with telescopes while looking at the stars. The *aberration* is a change of the expected light direction, which we will discuss shortly. At the Equator, Earth rotation adds an angular deviation of 0.32′, changing sign every 12 hours, to the aberration due to the motion of the Earth around the Sun, about 20.5′. In modern times, astronomers had found a number of additional proofs, but none was accessible to the man on the street.

Page 250

Furthermore, the measurements showing that the Earth is not a sphere, but *flattened* at the poles, confirmed the rotation of the Earth. Again, however, this eighteenth century measurement by Maupertuis*** is not accessible to everyday observation.

* 'And yet she moves' is the sentence falsely attributed to Galileo about the Earth. It is true, however, that in his trial he was forced to publicly retract the idea of a moving Earth to save his life (see the footnote on page 203).

** For the definition of angles see page 58 and for the definition of angle units see Appendix B.

*** Pierre Louis Moreau de Maupertuis (1698–1759), French physicist and mathematician. He was one

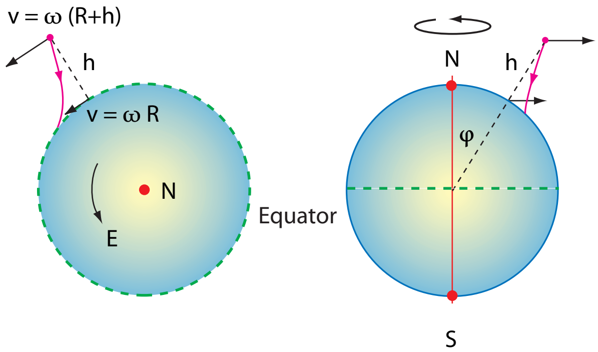


Figure 41 The deviations of free fall towards the east and towards the Equator due to the rotation of the Earth

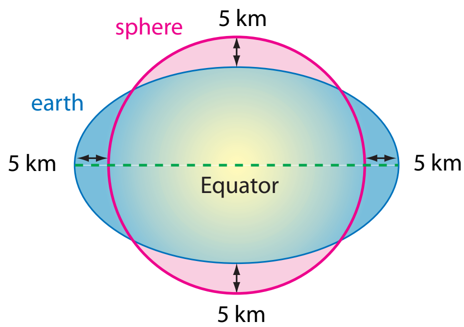


Figure 40 Earth's deviation from spherical shape due to its rotation

Then, in the years 1790 to 1792 in Bologna, Giovanni Battista Guglielmini (1763–1817) finally succeeded in measuring what Galileo and Newton had predicted to be the simplest proof for the Earth's rotation. On the Earth, objects do not fall vertically, but are slightly deviated to the east. This deviation appears because an object keeps the larger horizontal velocity it had at the height from which it started falling, as shown in Figure 41. Guglielmini's result was the first non-astronomical proof of the Earth's rotation. The experiments were repeated in 1802 by Johann Friedrich Benzenberg (1777–1846). Using metal balls which he dropped from the Michaelis tower in Hamburg – a height of 76 m – Benzenberg found that the deviation to the east was 9.6 mm. Can you confirm that the value measured by Benzenberg almost agrees with the assumption that the Earth turns once every 24 hours? (There is also a much smaller deviation towards the Equator, not measured by Guglielmini, Benzenberg or anybody after them up to this day; however, it completes the list of effects on free fall by the rotation of the Earth.) Both deviations are easily understood if we remember that falling objects describe an ellipse around the centre of the rotating Earth. The elliptical shape shows that the path of a thrown stone does not lie on a plane for an observer standing on Earth; for such an observer, the exact path thus cannot be drawn on a piece of paper.

Challenge 149 ny

In 1835, the French engineer and mathematician Gustave-Gaspard Coriolis (1792–1843), the same who also introduced the modern concepts of 'work' and of 'kinetic energy', found a closely related effect that nobody had noticed in everyday life up to then. An object travelling in a rotating background does not move on a straight line. If the rotation is anticlockwise, as is the case for the Earth on the northern hemisphere, the velocity of objects is slightly turned to the right, while its magnitude stays constant. This so-called

of the key figures in the quest for the principle of least action, which he named in this way. He was also founding president of the Berlin Academy of Sciences.

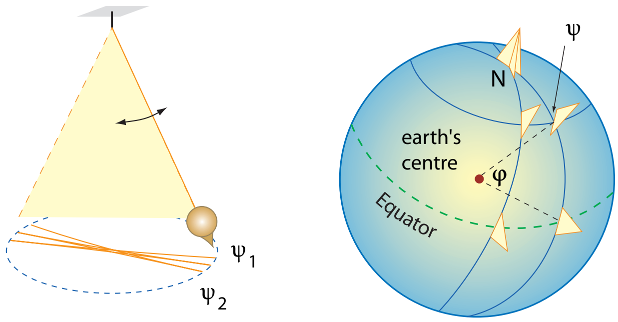


Figure 42 The turning motion of a pendulum showing the rotation of the Earth

Coriolis acceleration (or Coriolis force) is due to the change of distance to the rotation axis. Can you deduce the analytical expression for it, namely $\mathbf{a}_C = 2\omega \times \mathbf{v}$? Challenge 150 ny

The Coriolis acceleration determines the handedness of many large-scale phenomena with a spiral shape, such as the directions of cyclones and anticyclones in meteorology, the general wind patterns on Earth and the deflection of ocean currents and tides. Most beautifully, the Coriolis acceleration explains why icebergs do not follow the direction of the wind as they drift away from the polar caps. (Ref. 64) The Coriolis acceleration also plays a role in the flight of cannon balls (that was the original interest of Coriolis), in satellite launches, in the motion of sunspots and even in the motion of electrons in molecules. (Ref. 65) All these phenomena are of opposite sign on the northern and southern hemisphere and thus prove the rotation of the Earth. (In the First World War, many naval guns missed their targets in the southern hemisphere because the engineers had compensated them for the Coriolis effect on the northern hemisphere.)

(Ref. 66) Only in 1962, after several earlier attempts by other researchers, Asher Shapiro was the first to verify that the Coriolis effect has a tiny influence on the direction of the vortex formed by the water flowing out of a bathtub. Instead of a normal bathtub, he had to use a carefully designed experimental set-up, because contrary to an often-heard assertion, no such effect can be seen in real bathtubs. He succeeded only by carefully eliminating all disturbances from the system; for example, he waited 24 hours after the filling of the reservoir (and never actually stepped in or out of it!) in order to avoid any left-over motion of water that would disturb the effect, and built a carefully designed, completely rotationally-symmetric opening mechanism. Others have repeated the experiment in the southern hemisphere, confirming the result. (Ref. 66) In other words, the handedness of usual bathtub vortices is *not* caused by the rotation of the Earth, but results from the way the water starts to flow out. But let us go on with the story about Earth's rotation.

Finally, in 1851, the French physician turned physicist Jean Bernard Léon Foucault (b. 1819 Paris, d. 1868 Paris) performed an experiment that removed all doubts and rendered him world-famous practically overnight. He suspended a 67 m long pendulum* in the Panthéon in Paris and showed the astonished public that the direction of its swing changed over time, rotating slowly. To everybody with a few minutes of patience to watch

Challenge 151 d, Ref. 67
* Why was such a long pendulum necessary? Understanding the reasons allows one to repeat the experiment at home, using a pendulum as short as 70 cm, with the help of a few tricks.

the change of direction, the experiment proved that the Earth rotates. If the Earth did not rotate, the swing of the pendulum would always continue in the same direction. On a rotating Earth, the direction changes to the left, as shown in Figure 42, except if the pendulum is located at the Equator.* The time in which the orientation of the swing performs a full turn – the precession time – can be calculated. Study a pendulum swinging in the North–South direction and you will find that the precession time T_{Foucault} is given
Challenge 152 d by

$$T_{\text{Foucault}} = \frac{24\,\text{h}}{\sin\varphi} \qquad (27)$$

where φ is the latitude of the location of the pendulum, e.g. 0° at the Equator and 90° at the north pole. This formula is one of the most beautiful results of Galilean kinematics.*

Foucault was also the inventor and namer of the *gyroscope*. He built the device, shown in Figure 43, in 1852, one year after his pendulum. With it, he again demonstrated the rotation of the Earth. Once a gyroscope rotates, the axis stays fixed in space – but only when seen from far of the Earth. For an observer on Earth, the axis direction changes regularly with a period of 24 hours. Gyroscopes are now routinely used in ships and in aeroplanes to give the direction of north, because they are more precise and more reliable than magnetic compasses. In the most modern versions, one uses laser light running in circles instead of rotating masses.**

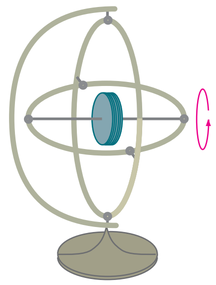


Figure 43 The gyroscope

In 1909, Roland von Eőtvős measured a simple effect: due to the rotation of the Earth, the weight of a object depends on the direction in which it moves. As a result, a balance in rotation around the vertical axis does not stay perfectly horizontal: the balance starts to oscillate slightly. Can you explain the origin of
Challenge 154 n the effect?

Ref. 68 In 1910, John Hagen published the results of an even simpler experiment, proposed by Louis Poinsot in 1851. Two masses are put on a horizontal bar that can turn around a vertical axis, a so-called *isotomeograph*. If the two masses are slowly moved towards the support, as shown in Figure 44, and if the friction is kept low enough, the bar rotates. Obviously, this would not happen if the Earth were not rotating. Can you explain the
Challenge 155 n observation? This little-known effect is also useful for winning bets among physicists.

In 1913, Arthur Compton showed that a closed tube filled with water and some small
Ref. 80 floating particles (or bubbles) can be used to show the rotation of the Earth. The device is called a *Compton tube* or *Compton wheel*. Compton showed that when a horizontal tube filled with water is rotated by 180°, something happens that allows one to prove that the Earth rotates. The experiment, shown in Figure 45, even allows measurement of the
Challenge 156 d latitude of the point where the experiment is made. Can you guess what happens?

* The discovery also shows how precision and genius go together. In fact, the first person to observe the effect was Vincenzo Viviani, a student of Galilei, as early as 1661! Indeed, Foucault had read about Viviani's work in the publications of the Academia dei Lincei. But it took Foucault's genius to connect the effect to the rotation of the Earth; nobody had done so before him.

* The calculation of the period of Foucault's pendulum assumed that the precession rate is constant during a rotation. This is only an approximation (though usually an good one).

Challenge 153 n ** Can you guess how rotation is detected in this case?

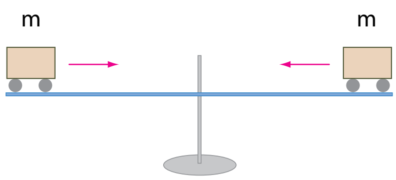


Figure 44 Showing the rotation of the Earth through the rotation of an axis

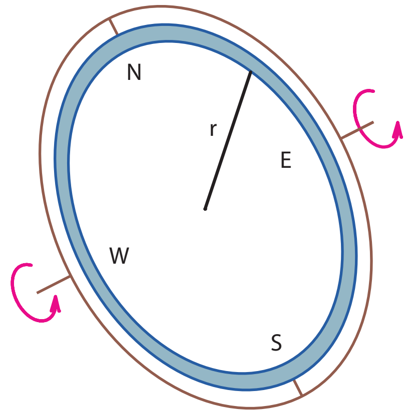


Figure 45 Demonstrating the rotation of the Earth with water

In 1925, Albert Michelson* and his collaborators in Illinois constructed a vacuum interferometer with the incredible perimeter of 1.9 km. Interferometers produce bright and Page 678 dark fringes of light; the position of the fringes depends on the speed at which the interferometers rotates. The fringe shift is due to an effect first measured in 1913 by the French physicist Georges Sagnac: the rotation of a complete ring interferometer with angular Challenge 157 ny frequency (vector) $\mathbf{\Omega}$ produces a fringe shift of angular phase $\Delta\varphi$ given by

$$\Delta\varphi = \frac{8\pi\, \mathbf{\Omega}\, \mathbf{A}}{c\, \lambda} \qquad (28)$$

where $\mathbf{A}$ is the area (vector) enclosed by the two interfering light rays, λ their wavelength and c the speed of light. The effect is now called the *Sagnac effect*, even though it had Ref. 69 been predicted already 20 years earlier by Oliver Lodge.* Michelson and his team found a fringe shift with a period of 24 hours and of exactly the magnitude predicted by the rotation of the Earth. Modern high precision versions use ring lasers with areas of only a few square metres, but are able to measure variations of the rotation rates of the Earth of less than one part per million. Indeed, over the course of a year the length of a day varies irregularly by a few milliseconds, mostly due to influences from the Sun or the Moon, Ref. 70 due to weather changes and due to hot magma flows deep inside the Earth.** But also earthquakes, the el Ninño effect in the climate and the filling of large water dams have effects on the rotation of the Earth. All these effects can be studied with such precision

* Albert Abraham Michelson (b. 1852 Strelno, d. 1931 Pasadena) Prussian-Polish-US-American physicist, obsessed by the precise measurement of the speed of light, received the Nobel Prize in physics in 1907.

* Oliver Lodge (1851–1940) was a British physicist who studied electromagnetic waves and tried to communicate with the dead. A strange but influential figure, his ideas are often cited when fun needs to be made of physicists; for example, he was one of those physicists who believed that at the end of the nineteenth century physics was complete.

** The growth of leaves on trees and the consequent change in the Earth's moment of inertia, studied already in 1916 by Harold Jeffreys, is too small to be seen, so far.

interferometers; these apparatus can also be used for research into the motion of the soil due to lunar tides or earthquakes, and for checks on the theory of relativity.

In summary, observations show that the Earth surface rotates at 463 m/s at the Equator, a larger value than that of the speed of sound in air – about 340 m/s in usual conditions – and that we are in fact *whirling* through the universe.

How does the Earth rotate?

Is the rotation of the Earth *constant* over geological time scales? That is a hard question. If you find a method leading to an answer, publish it! (The same is true for the question whether the length of the year is constant.) Only a few methods are known, as we will find out shortly. Ref. 71

The rotation of the Earth is not even constant during human lifespans. It varies by a few parts in 10^8. In particular, on a 'secular' time scale, the length of the day increases by about 1 to 2 ms per century, mainly because of the friction by the Moon and the melting of the ice caps at the poles. This was deduced by studying historical astronomical observations by the ancient Babylonian and Arab astronomers. Additional 'decadic' changes have an amplitude of 4 or 5 ms and are due to the motion of the liquid part of the Earth's core. Ref. 72

The seasonal and biannual changes of the length of the day – with an amplitude of 0.4 ms over six months, another 0.5 ms over the year, and 0.08 ms over 24 to 26 months – are mainly due to the effects of the atmosphere. In the 1950s the availability of precision measurements showed that there is even a 14 and 28 day period with an amplitude of 0.2 ms, due to the Moon. In the 1970s, when wind oscillations with a length scale of about 50 days were discovered, they were also found in the length of the day, with an amplitude of about 0.25 ms. However, these last variations are quite irregular.

But why does the Earth rotate at all? The rotation derives from the rotating gas cloud at the origin of the solar system. This connection explains that the Sun and all planets, except one, turn around themselves in the same direction, and that they also all turn around the Sun in that same direction. But the complete story is outside the scope of this text. Ref. 73

The rotation around its axis is not the only motion of the Earth; it performs other motions as well. This was known already long ago. In 128 BCE, the Greek astronomer Hipparchos discovered what is today called the *(equinoctial) precession*. He compared a measurement he made himself with another made 169 years before. Hipparchos found that the Earth's axis points to different stars at different times. He concluded that the sky was moving. Today we prefer to say that the axis of the Earth is moving. During a period of 25 800 years the axis draws a cone with an opening angle of 23.5°. This motion, shown in Figure 46, is generated by the tidal forces of the Moon and the Sun on the equatorial bulge of the Earth resulting form its flattening. The Sun and the Moon try to put the axis of the Earth at a right angle to the Earth's path; this torque leads to the precession of the Earth's axis. (The same effect appears for any spinning top or in the experiment with the suspended wheel shown on page 159.)

In addition, the axis of the Earth is not even fixed compared to the Earth's surface. In 1884, by measuring the exact angle above the horizon of the celestial North Pole, Friedrich Küstner (1856–1936) found that the axis of the Earth *moves* with respect to the Earth's crust, as Bessel had suggested 40 years earlier. As a consequence of Küstner's discov-

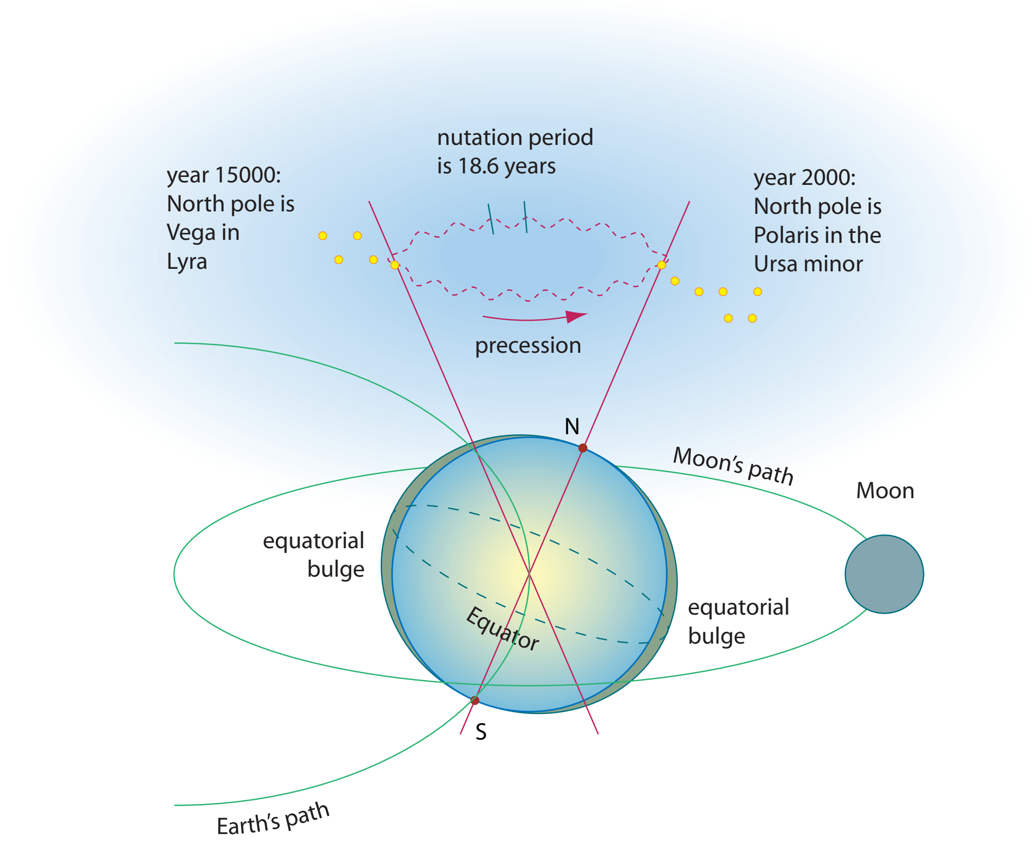


Figure 46 The precession and the nutation of the Earth's axis

ery, the International Latitude Service was created. The *polar motion* Küstner discovered turned out to consist of three components: a small linear drift – not yet understood – a yearly elliptical motion due to seasonal changes of the air and water masses, and a circular motion* with a period of about 1.2 years due to fluctuations in the pressure at the bottom of the oceans. In practice, the North Pole moves with an amplitude of 15 m around an average central position. Ref. 74

In 1912, the German meteorologist and geophysicist Alfred Wegener (1880–1930) discovered an even larger effect. After studying the shapes of the continental shelves and the geological layers on both sides of the Atlantic, he conjectured that the continents *move*, and that they are all fragments of a single continent that broke up 200 million years ago.*

* The circular motion, a wobble, was predicted by the great Swiss mathematician Leonhard Euler (1707–1783). Using this prediction and Küstner's data, in 1891 Seth Chandler claimed to be the discoverer of the circular component.

* In this old continent, called Gondwanaland, there was a huge river that flew westwards from the Chad to Guayaquil in Ecuador. After the continent split up, this river still flowed to the west. When the Andes appeared, the water was blocked, and many millions of years later, it flowed back. Today, the river still flows eastwards and is called the Amazonas.

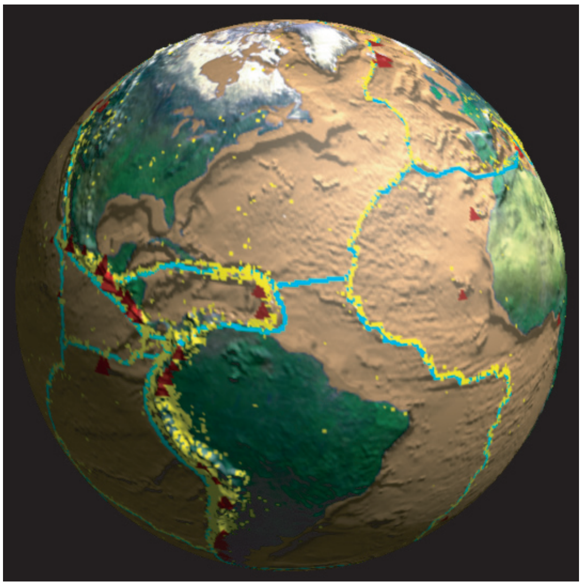


Figure 47 The continental plates as objects of tectonic motion

Even though derided at first across the world, his discoveries were correct. Modern satellite measurements, shown in Figure 47, confirm this model. For example, the American continent moves away from the European continent by about 10 mm every year. There are also speculations that this velocity may have been much higher for certain periods in the past. The way to check this is to look at magnetization of sedimental rocks. At present, this is still a hot topic of research. Following the modern version of the model, called *plate tectonics*, the continents (with a density of $2.7 \cdot 10^3$ kg/m^3) float on the fluid mantle of the Earth (with a density of $3.1 \cdot 10^3$ kg/m^3) like pieces of cork on water, and the convection inside the mantle provides the driving mechanism for the motion.

Page 100, page 554
Ref. 75

Does the Earth move?

The centre of the Earth is not at rest in the universe. In the third century BCE Aristarchos of Samos had already maintained that the Earth turns around the Sun. However, a fundamental difficulty of the heliocentric system is that the stars look the same all year long. How can this be, if the Earth travels around the Sun? The distance between the Earth and the Sun has been known since the seventeenth century, but it was only in 1837 that Friedrich Wilhelm Bessel* became the first to observe the *parallax* of a star. This was a result of extremely careful measurements and complex calculations: he discovered the *Bessel functions* in order to realize it. He was able to find a star, 61 Cygni, whose apparent position changed with the month of the year. Seen over the whole year, the star describes a small ellipse on the sky, with an opening of 0.588″ (this is the modern value). After

* Friedrich Wilhelm Bessel (1784–1846), Westphalian astronomer who left a successful business career to dedicate his life to the stars, and became the foremost astronomer of his time.

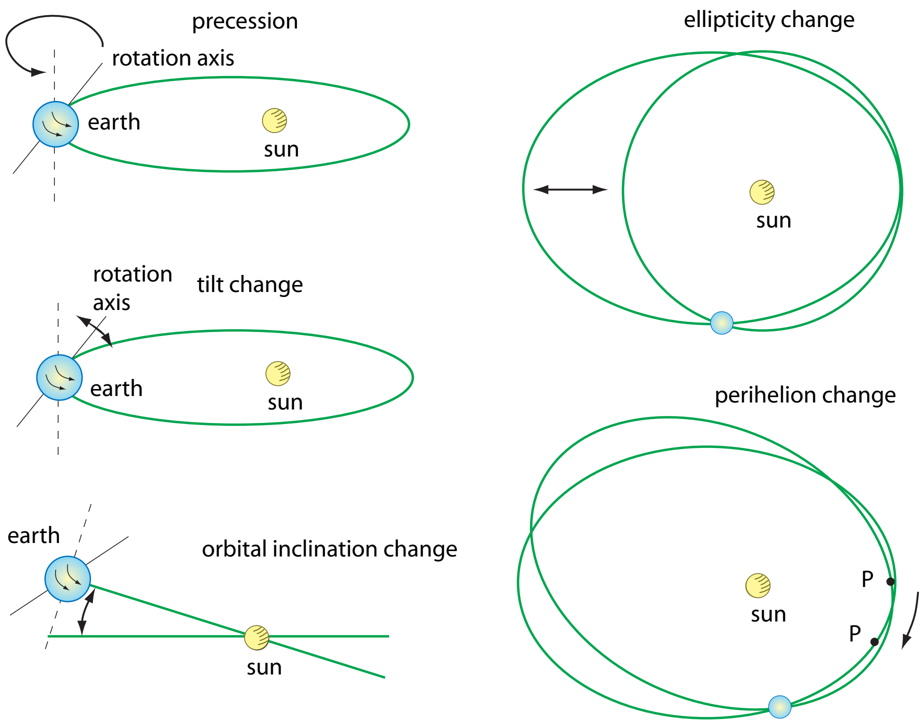


Figure 48 Changes in the Earth's motion around the Sun

carefully eliminating all other possible explanations, he deduced that the change of position was due to the motion of the Earth around the Sun, and from the size of the ellipse he determined the distance to the star to be 105 Pm, or 11.1 light years.
Challenge 158 n

Bessel had thus managed for the first time to measure the distance of a star. By doing so he also proved that the Earth is not fixed with respect to the stars in the sky and that the Earth indeed revolves around the Sun. The motion itself was not a surprise. It confirmed the result of the mentioned aberration of light, discovered in 1728 by James Bradley and to be discussed shortly; the Earth moves around the Sun.
Page 250

With the improvement of telescopes, other motions of the Earth were discovered. In 1748, James Bradley announced that there is a small regular *change* of the precession, which he called *nutation*, with a period of 18.6 years and an angular amplitude of 19.2″. Nutation appears because the plane of the Moon's orbit around the Earth is not exactly the same as the plane of the Earth's orbit around the Sun. Are you able to confirm that this situation can produce nutation?
Challenge 159 ny

Astronomers also discovered that the 23.5° tilt – or *obliquity* – of the Earth's axis, the angle between its intrinsic and its orbital angular momentum, actually changes from 22.1° to 24.5° with a period of 41 000 years. This motion is due to the attraction of the Sun and the deviations of the Earth from a spherical shape. During the Second World War, in 1941, the Serbian astronomer Milutin Milankovitch (1879–1958) retreated into solitude and studied the consequences. In his studies he realized that this 41 000 year period of

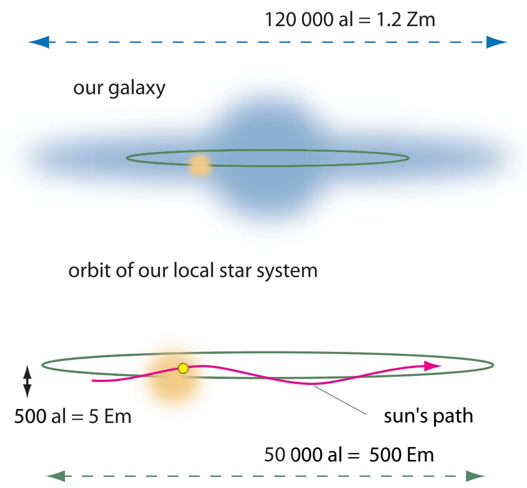


Figure 49 The motion of the Sun around the galaxy

the tilt, together with an average period of 22 000 years due to precession,* gives rise to the more than 20 *ice ages* in the last 2 million years. This happens through stronger or weaker irradiation of the poles by the Sun. The changing amounts of melted ice then lead to changes in average temperature. The last ice age had is peak about 20 000 years ago and finished around 10 000 years ago; the next is still far away. A spectacular confirmation of the ice age cycles, in addition to the many geological proofs, came through measurements of oxygen isotope ratios in sea sediments, which allow the average temperature over the past million years to be tracked. Ref. 76

The Earth's orbit also changes its *eccentricity* with time, from completely circular to slightly oval and back. However, this happens in very complex ways, not with periodic regularity, and is due to the influence of the large planets of the solar system on the Earth's orbit. The typical time scale is 100 000 to 125 000 years.

In addition, the Earth's orbit changes in *inclination* with respect to the orbits of the other planets; this seems to happen regularly every 100 000 years. In this period the inclination changes from +2.5° to −2.5° and back.

Even the direction in which the ellipse points changes with time. This so-called *perihelion shift* is due in large part to the influence of the other planets; a small remaining part will be important in the chapter on general relativity. It was the first piece of data confirming the theory.

Obviously, the length of the year also changes with time. The measured variations are of the order of a few parts in 10^{11} or about 1 ms per year. However, the knowledge of these changes and of their origins is much less detailed than for the changes in the Earth's rotation.

The next step is to ask whether the Sun itself moves. Indeed it does. Locally, it moves with a speed of 19.4 km/s towards the constellation of Hercules. This was shown by Wil-

* In fact, the 25 800 year precession leads to three insolation periods, of 23 700, 22 400 and 19 000 years, due to the interaction between precession and perihelion shift.

liam Herschel in 1783. But globally, the motion is even more interesting. The diameter of the galaxy is at least 100 000 light years, and we are located 26 000 light years from the centre. (This is known since 1918; the centre of the galaxy is located in the direction of Sagittarius.) At our position, the galaxy is 1 300 light years thick; presently, we are 68 light years 'above' the centre plane. The Sun, and with it the solar system, takes about 225 million years to turn once around the galactic centre, its orbital velocity being around 220 km/s. It seems that the Sun will continue moving away from the galaxy plane until it is about 250 light years above the plane, and then move back, as shown in Figure 49. The oscillation period is estimated to be around 60 million years, and has been suggested as mechanism for the mass extinctions of animal life on Earth, possibly because some gas cloud is encountered on the way. The issue is still a hot topic of research.

Ref. 77

We turn around the galaxy centre because the formation of galaxies, like that of solar systems, always happens in a whirl. By the way, are you able to confirm by your own observation that our galaxy itself rotates?

Challenge 160 n

Finally, we can ask whether the galaxy itself moves. Its motion can indeed be observed because it is possible to give a value for the motion of the Sun through the universe, defining it as the motion against the background radiation. This value has been measured to be 370 km/s. (The velocity of the *Earth* through the background radiation of course depends on the season.) This value is a combination of the motion of the Sun around the galaxy centre and of the motion of the galaxy itself. This latter motion is due to the gravitational attraction of the other, nearby galaxies in our local group of galaxies.*

Ref. 78

In summary, the Earth really moves, and does so in rather complex ways. As Henri Poincaré would say, if we are in a given spot today, say the Panthéon in Paris, and come back to the same spot tomorrow at the the same time, we are in fact 31 million kilometres away. This state of affairs would make time travel extremely difficult even if it were possible (which it is not); whenever you would go back to the past, you would have to get exactly to the old spot!

Is rotation relative?

When we turn rapidly, our arms lift. Why does this happen? How can our body detect whether we rotate or not? There are two possible answers. The first approach, promoted by Newton, is to say that there is an absolute space; whenever we rotate against this space, the system reacts. The other answer is to note that whenever the arms lift, also the stars rotate, and in exactly the same manner. In other words, our body detects rotation because we move against the average mass distribution in space.

The most cited discussion of this question is due to Newton. Instead of arms, he explored the water in a rotating bucket. As usual for philosophical issues, Newton's answer was guided by the mysticism triggered by his father's absence. Newton sees absolute space as a religious concept and is not even able to conceive the second alternative. Newton thus sees rotation as an absolute concept. Most modern scientist have fewer problems and more common sense than Newton; as a result, today's consensus is that rotation effects are due to the mass distribution in the universe: rotation is relative. However, we have

* This is roughly the end of the ladder. Note that the expansion of the universe, to be studied later on, produces no motion.

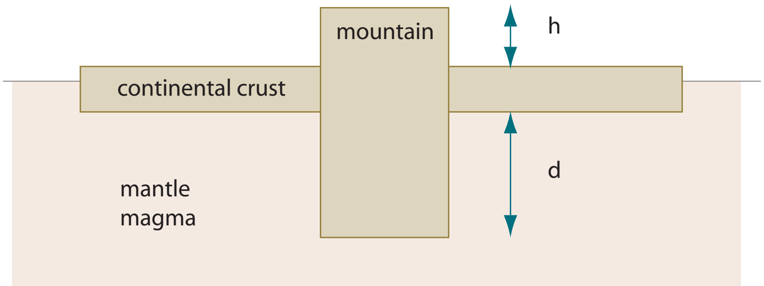


Figure 51 A simple model for continents and mountains

to be honest; the question cannot be settled by Galilean physics. We will need general relativity to do so.

Curiosities and fun challenges about everyday motion

> It is a mathematical fact that the casting of this pebble from my hand alters the centre of gravity of the universse.
>
> Thomas Carlyle,* *Sartor Resartus III.*

Here are a few facts to ponder about motion.

▪ A train starts to travel at a constant speed of 10 m/s between two cities A and B, with a distance of 36 km. The train will take one hour for the journey. At the same time as the train, a fast dove starts to fly from A to B, at 20 m/s. Being faster than the train, the dove arrives at B first. The dove then flies back to the train; when it meets the train, it turns again to the city B, and goes on flying back and forward until the train has arrived at B.

Challenge 162 e What distance did the dove cover?

▪ A good bathroom scale, used to determine the weight of objects, does not show a constant weight when

Challenge 163 n you step on it and stay motionless. Why not?

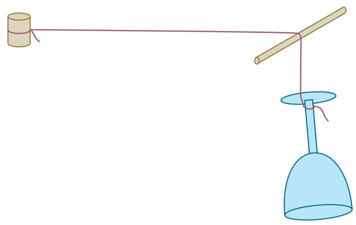


Figure 50 Is it safe to let the cork go?

▪ A cork is attached to a thin string of a metre in length. The string is passed over a long rod held horizontally, and a wine glass is attached at the other end. If you let go the cork in Figure 50, nothing breaks. Why

Challenge 164 n not? And what happens?

▪ In 1901, Duncan MacDougalls, a medical doctor, measured the weight of dying people, in the hope to see whether death leads to a mass change. He found a sudden change of about 20 g at the moment of death, with large variations from person to person, and attributed it to the

Challenge 165 ny soul. Is this explanation satisfactory? (If you know a better one, publish it!)

▪ The Earth's crust is less dense (2.7 kg/l) than the Earth's mantle (3.1 kg/l) and swims on it. As a result, the lighter crust below a mountain ridge must be much deeper than below a plane. If a mountain is 1 km higher than the plane, how much deeper must the

Challenge 166 n crust be below it? The block model works fairly well; first, it explains why near moun-

Challenge 161 n * Thomas Carlyle (1797–1881). Do you agree with the quotation?

Figure 52 A well-known toy

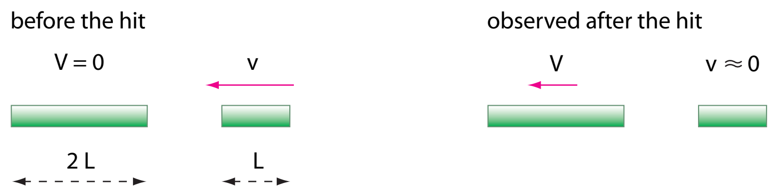


Figure 53 An elastic collision that seems not to obey energy conservation

tains measurements of the deviation of free fall from the vertical line lead to so much smaller values than those expected without a deep section. Later, sound measurements have confirmed directly that the continental crust is indeed thicker below mountains.

- Put a vertical pencil on a paper on the border of a table. How can you pull out the paper without letting the pencil fall? Challenge 167 e

- Take a pile of coins. One can pull out the coins, starting with the one at the bottom, by shooting another coin over the table surface. The method also helps to visualize two-dimensional momentum conservation. Challenge 168 e

- In early 2004, two men and a woman earned 1.2 million British pounds in a single evening in a London casino. They did so by applying the formulas of Galilean mechanics. They used the method pioneered by various physicists in the 1950s who built various small computers that could predict the outcome of a roulette ball from the initial velocity imparted by the croupier. In the case in Britain, the group added a laser scanner to a smart phone that measured the path of a roulette ball and predicted the numbers where it would arrive. In this way, they increased the odds from 1 in 37 to about 1 in 6. After six months of investigations, Scotland Yard ruled that they could keep the money they won. Ref. 79

- Is a return flight by plane – from a point A to B and back to A – faster if the wind blows or not? Challenge 169 e

- The toy of Figure 52 shows interesting behaviour: when a number of spheres are lifted and made to hit the resting ones, the same number of spheres detach on the other side, whereas the previously dropped spheres remain motionless. At first sight, all this seems to follow from energy and momentum conservation. However, energy and momentum conservation only provide two equations, which are not sufficient to explain or determine the behaviour of five spheres. Why then do the spheres behave in this way? And why do they swing all in phase for long times? Challenge 170 d

- A surprising effect is used in home tools such as hammer drills. We remember that when a small ball hits elastically a large one at rest, both balls move after the hit, and the small one obviously moves faster than the large one. Ref. 81

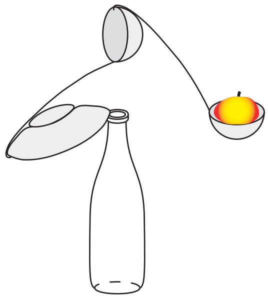


Figure 54 The centre of mass defines stability

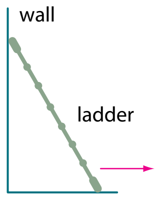


Figure 55 How does the ladder fall?

Despite this result, when a short cylinder hits a long one of the same diameter and material, but with a length which is some *integer* multiple of that of the short one, something strange happens. After the hit, the small cylinder remains almost at rest, whereas the large one moves, as shown in Figure 53. Conservation of momentum seems not to hold at all in this case. (In fact this is the reason that demonstrations with elastic collisions in schools are always performed with spheres.) What is the origin of this effect? Challenge 171 d

▪ Does a wall get a stronger jolt when it is hit by a ball rebounding from it or when it is hit by a ball that remains stuck to it? Challenge 172 n

▪ Housewives know how to extract a cork of a wine bottle using a cloth. Can you imagine how? They also know how to extract the cork with the cloth if the cork has fallen inside the bottle. How? Challenge 173 n

▪ The sliding ladder problem, shown schematically in Figure 55, asks for the detailed motion of the ladder over time. The problem is more difficult than it looks, even if friction is not taken into account. Can you say whether the lower end always touches the floor? Challenge 174 ny

▪ A common fly on the stern of a 30 000 ton ship of 100 m length tilts it by less than the diameter of an atom. Today, distances that small are easily measured. Can you think of at least two methods, one of which should not cost more than 2000 Euro? Ref. 82 Challenge 175 n

▪ The level of acceleration a human can survive depends on the duration one is subjected to it. For a tenth of a second, $30\,g = 300\,\mathrm{m/s^2}$, as generated by ejector seats in aeroplanes, is acceptable. (It seems that the record acceleration a human survived is about $80\,g = 800\,\mathrm{m/s^2}$.) But as a rule of thumb it is said that accelerations of $15\,g = 150\,\mathrm{m/s^2}$ or more are fatal.

▪ The highest *microscopic* accelerations are observed in particle collisions, where one gets values up to $10^{35}\,\mathrm{m/s^2}$. The highest *macroscopic* accelerations are probably found in the collapsing interiors of *supernovae*, exploding stars which can be so bright as to be visible in the sky even during the daytime. A candidate on Earth is the interior of collapsing bubbles in liquids, a process called *cavitation*. Cavitation often produces light, an effect discovered by Frenzel and Schulte in 1934 and called *sonoluminescence*. (See Figure 56.) It appears most prominently when air bubbles in water are expanded and contracted by underwater loudspeakers at around 30 kHz and allows precise measurements of bubble motion. At a certain threshold intensity, the bubble radius changes at 1500 m/s in as little no figure yet

Figure 56 Observation of sonoluminescence and a diagram of the experimental set-up

Ref. 83 as a few μm, giving an acceleration of several 10^{11} m/s^2.

▪ If a gun located at the Equator shoots a bullet in the vertical direction, where does the bullet fall back? Challenge 176 n

▪ Why are most rocket launch sites as near as possible to the Equator? Challenge 177 n

▪ Is travelling through interplanetary space healthy? People often fantasize about long trips through the cosmos. Experiments have shown that on trips of long duration, cosmic radiation, bone weakening and muscle degeneration are the biggest dangers. Many medical experts question the viability of space travel lasting longer than a couple of years. Other dangers are rapid sunburn, at least near the Sun, and exposure to the vacuum. So Ref. 84 far only one man has experienced vacuum without protection. He lost consciousness after 14 seconds, but survived unharmed.

▪ How does the kinetic energy of a rifle bullet compare to that of a running man? Challenge 178 n

▪ In which direction does a flame lean if it burns inside a jar on a rotating turntable? Challenge 179 n

▪ A ping-pong ball is attached with a string to a stone, and the whole is put under water in a jar. The set-up is shown in Figure 57. Now the jar is accelerated horizontally. In which direction does the ball move? What do you deduce for a jar at rest? Challenge 180 n

▪ What happens to the size of an egg when one places it into a jar of vinegar for a few days? Challenge 181 n

▪ Does centrifugal acceleration exist? Most university students go through the shock of meeting a teacher who says that it doesn't because it is a 'fictitious' quantity, in the face of what one experiences every day in a car when driving around a bend. Simply ask the teacher who denies it to define 'existence'. (The definition physicists usually use is given in the Intermezzo following this chapter.) Page 628 Then check whether the definition applies to the term and make up your own mind. Challenge 182 ny

▪ Rotation holds a surprise for everybody studying it carefully. Angular momentum is a quantity with a magnitude and a direction. However, it is not a vector, as any mirror shows. The angular momentum of a body circling in a plane parallel to a mirror behaves differently from a usual arrow: its mirror image is not reflected if it points towards the mirror! You can easily check this by yourself. Challenge 183 e For this reason, angular momentum is called a *pseudovector*. The fact has no important consequences in classical physics; but we have to keep it in mind for later occasions.

▪ What is the best way to transport full coffee or tea cups while at the same time avoiding spilling any precious liquid? Challenge 184 n

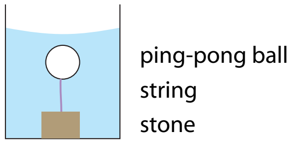


Figure 57 How does the ball move when the jar is accelerated?

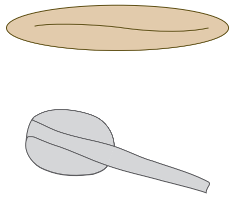


Figure 58 The famous Celtic stone and a version made with a spoon

▪ The Moon recedes from the Earth by 3.8 cm a year, due to friction. Can you find the mechanism responsible?

Challenge 185 ny

▪ What is the amplitude of a pendulum oscillating in such a way that the absolute value of its acceleration at the lowest point and at the return point are equal?

Challenge 186 ny

▪ Can you confirm that the value of the acceleration of a drop of water falling through vapour is $g/7$?

Challenge 187 ny

▪ What are earthquakes? *Earthquakes* are large examples of the same process that makes doors squeak. The continental plates correspond to the metal surfaces in the joints of the door.

Earthquakes can be described as energy sources. The Richter scale is a direct measure of this energy. The *Richter magnitude* M_s of an earthquake, a pure number, is defined from its energy E via

$$M_s = \frac{\log(E/1\,\mathrm{J}) - 4.8}{1.5} \,. \qquad (29)$$

The strange numbers in the expression have been chosen to put the earthquake values as near as possible to the older, qualitative Mercalli scale (now called EMS98) that classifies the intensity of earthquakes. However, this is not fully possible; the most sensitive instruments today detect earthquakes with magnitudes of −3; the highest value every measured was a Richter magnitude of 10, in Chile in 1960. Magnitudes above 12 are probably impossible. (Can you show why?)

Challenge 188 n

▪ Figure 58 shows the so-called *Celtic wiggle stone*, a stone that starts rotating on a plane surface when it is put into oscillation. The size can vary between a few centimetres and a few metres. Simply by bending a spoon one can realize a primitive form of this strange device, if the bend is not completely symmetrical. The rotation is always in the same direction. If the stone is put into rotation in the wrong direction, after a while it stops and starts rotating in the other sense! Can you explain the effect?

Ref. 81

Challenge 189 d

▪ What is the motion of the point below the Sun on a map of the Earth during one day, and day after day?

Challenge 190 ny

▪ The moment of inertia of a body does depend on the shape of the body; usually, angular momentum and the angular velocity do not point in the same direction. Can you confirm this with an example?

Challenge 191 n

▪ Can it happen that a satellite dish for geostationary TV satellites focuses the sunshine

Challenge 192 n onto the receiver?

▪ Why is it difficult to fire a rocket from an aeroplane in the direction opposite to the motion of the plane? Challenge 193 n

▪ You have two hollow spheres: they have the same weight, the same size and painted the same colour. One is made of copper, the other of aluminium. Obviously, they fall with the same speed and acceleration. What happens if they both roll down a tilted plane? Challenge 194 ny

▪ An ape hangs on a rope. The rope hangs over a wheel and is attached to a mass of equal weight hanging down on the other side, as shown in Figure 59. The rope is massless, the wheel massless and frictionless. What happens when the ape climbs the rope? Challenge 195 n

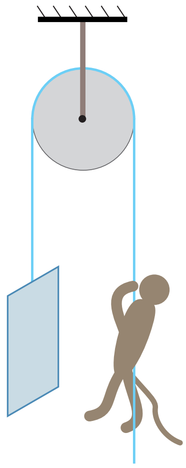


Figure 59 How does the ape move?

▪ What is the shape of a rope when rope jumping? Challenge 196 ny

▪ How can you determine the speed of a rifle bullet only with a scale and a metre stick? Challenge 197 n

▪ Why does a gun make a hole in a door but cannot push it open, in exact contrast to what a finger does? Challenge 198 e

▪ Can a water skier move with a higher speed than the boat pulling him? Challenge 199 n

▪ Take two cans of the same size and weight, one full of ravioli and one full of peas. Which one rolls faster on an inclined plane? Challenge 200 ny

▪ What is the moment of inertia of a homogeneous sphere? Challenge 201 n

▪ The moment of inertia is determined by the values of its three principal axes. These are all equal for a sphere and for a cube. Does it mean that it is impossible to distinguish a sphere from a cube by their inertial behaviour? Challenge 202 ny

▪ You might know the 'Dynabee', a hand-held gyroscopic device that can be accelerated to high speed by proper movements of the hand. How does it work? Challenge 203 d

▪ Is it true that the Moon in the first quarter in the northern hemisphere looks like the Moon in the last quarter in the southern hemisphere? Challenge 204 n

▪ An impressive confirmation that the Earth is round can be seen at sunset, if one turns, against usual habits, the back to the Sun. On the eastern sky one can see the impressive rise of the Earth's shadow. (In fact, more precise investigations show that it is not the shadow of the Earth alone, but the shadow of its ionosphere.) One can admire a vast shadow rising over the whole horizon, clearly having the shape of a segment of a huge circle.

▪ How would Figure 60 look if taken at the Equator? Challenge 205 n

▪ Since the Earth is round, there are many ways to drive from one point on the Earth to another along a circle segment. This has interesting consequences for volley balls and for looking at women. Take a volleyball and look at its air inlet. If you want to move the inlet to a different position with a simple rotation, you can choose the rotation axis in may different ways. Can you confirm this? In other words, when we look in a given direction and then want to change to another, the eye can realize this change in different ways. The option chosen by the human eye had already been studied by medical scientists in the eighteenth century. It is called *Listing's 'law'*.* It states that all axes that nature chooses lie Challenge 206 e

* If you are interested in learning in more detail how nature and the eye cope with the complexities of three dimensions, see the http://schorlab.berkeley.edu/vilis/whatisLL.htm and http://www.med.uwo.ca/

Figure 60 A long exposure of the stars at night

Challenge 207 n in one plane. Can you imagine its position in space? Men have a deep interest that this mechanism is being followed; if not, on the beach, when men look at one woman after another, the muscles moving the eyes could get knotted up.

Legs or wheels? – Again

The acceleration and deceleration of standard wheel-driven cars is never much higher than about $1\,g = 9.8\,\mathrm{m/s^2}$, the acceleration due to gravity on our planet. Higher accelerations are achieved by motor bikes and racing cars through the use of suspensions that divert weight to the axes and by the use of spoilers, so that the car is pushed downwards with more than its own weight. Modern spoilers are so efficient in pushing a car towards the track that racing cars could race on the roof of a tunnel without falling down.

Through the use of special tyres these downwards forces are transformed into grip; modern racing tyres allow forward, backward and sideways accelerations (necessary for speed increase, for braking and for turning corners) of about 1.1 to 1.3 times the load. Engineers once believed that a factor 1 was the theoretical limit and this limit is still sometimes found in textbooks; but advances in tyre technology, mostly by making clever use of interlocking between the tyre and the road surface as in a gear mechanism, have allowed engineers to achieve these higher values. The highest accelerations, around $4\,g$,

physiology/courses/LLConsequencesWeb/ListingsLaw/perceptual2.htm websites.

Figure 61 A basilisk lizard (*Basiliscus basiliscus*) running on water, showing how the propulsing leg pushes into the water

are achieved when part of the tyre melts and glues to the surface. Special tyres designed to make this happen are used for dragsters, but high performance radio-controlled model cars also achieve such values.

How do all these efforts compare to legs? High jump athletes can achieve peak accelerations of about 2 to 4 g, cheetahs over 3 g, bushbabies up to 13 g, locusts about 18 g, and fleas have been measured to accelerate about 135 g. The maximum acceleration known for animals is that of click beetles, a small insect able to accelerate at over 2000 $\mathrm{m/s^2}$ = 200 g, about the same as an airgun pellet when fired. Legs are thus definitively more efficient accelerating devices than wheels – a cheetah easily beats any car or motorbike – and evolution developed legs, instead of wheels, to improve the chances of an animal in danger to get to safety. In short, legs *outperform* wheels.

Ref. 85

Challenge 208 n

There are other reasons to use legs instead of wheels. (Can you name some?) For example, legs, in contrast to wheels, allow walking on water. Most famous for this ability is the *basilisk*,* a lizard living in Central America. This reptile is about 50 cm long and has a mass of about 90 g. It looks like a miniature Tyrannosaurus Rex and is able to run over water surfaces on its hind legs. The motion has been studied in detail with high-speed cameras and by measurements using aluminium models of the animal's feet. The experiments show that the feet slapping on the water provides only 25 % of the force necessary to run above water; the other 75 % is provided by a pocket of compressed air that the basilisks create between their feet and the water once the feet are inside the water. In fact, basilisks mainly walk on air.** It was calculated that a human is also able to walk on water, provided his feet hit the water with a speed of 100 km/h using the simultaneous physical power of 15 sprinters. Quite a feat for all those who ever did so.

Ref. 86

There is a second method of walk and running on water; this second method even allows its users one to remain immobile on top of the water surface. This is what water

* In the Middle Ages, the term 'basilisk' referred to a mythical monster supposed to appear shortly before the end of the world. Today, it is a small reptile in the Americas.

Ref. 87

** Both effects used by basilisks are also found in fast canoeing.

striders, insects of the family *Gerridae* with a overall length of up to 15 mm, are able to do (together with several species of spiders). Like all insects, the water strider has six legs (spiders have eight). The water strider uses the back and front legs to hover over the surface, helped by thousands of tiny hairs attached to its body. The hairs, together with the surface tension of water, prevent the strider from getting wet. If you put shampoo into the water, the water strider sinks and cannot move any more. The water strider uses its large middle legs as oars to advance over the surface, reaching speeds of up to 1 m/s doing so. In short, water striders actually row over water.

Figure 62 A water strider (© Charles Lewallen)

Legs pose many interesting problems. Engineers know that a staircase is comfortable to walk only if for each step the length plus *twice* the height is about 63 ± 2 cm. This is the so-called *staircase formula*. Why does it hold?

Challenge 209 ny

All animals have an *even* number of legs. Do you know an exception? Why not? In fact, one can argue that no animal has less than four legs. Why is this the case?

Challenge 210 n

On the other hand, all animals with two legs have the legs side by side, whereas systems with wheels have them one behind the other. Why is this not the other way round?

Challenge 211 ny

But let us continue with the study of motion transmitted over distance, without the use of any contact at all.

Dynamics due to gravitation

Caddi come corpo morto cade.
Dante, *Inferno*, c. V, v. 142.*

The first and main contact-free method to generate motion we discover in our environment is *height*. Waterfalls, snow, rain and falling apples all rely on it. It was one of the fundamental discoveries of physics that height has this property because there is an interaction between every body and the Earth. *Gravitation* produces an acceleration along the line connecting the centres of gravity of the two bodies. Note that in order to make this statement, it is necessary to realize that the Earth is a body in the same way as a stone or the Moon, that this body is finite and that therefore it has a centre and a mass. Today, these statements are common knowledge, but they are by no means evident from everyday personal experience.**

* 'I fell like dead bodies fall.' Dante Alighieri (1265, Firenze–1321, Ravenna), the powerful Italian poet.

** In several myths about the creation or the organization of the world, such as the biblical one or the Indian one, the Earth is not an object, but an imprecisely defined entity, such as an island floating or surrounded by water with unclear boundaries and unclear suspension method. Are you able to convince a friend that the Earth is round and not flat? Can you find another argument apart from the roundness of the Earth's shadow when it is visible on the Moon?

Challenge 212 n

Ref. 88

A famous crook, Robert Peary, claimed to have reached the North Pole in 1909. (In fact, Roald Amundsen reached the both the South and the North Pole first.) Peary claimed to have taken a picture there, but that picture, which went round the world, turned out to be the proof that he had not been there. Can you imagine how?

Challenge 213 n

By the way, if the Earth is round, the top of two buildings is further apart than their base. Can this effect be measured?

Challenge 214 n

How does gravitation change when two bodies are far apart? The experts for distant objects are the astronomers. Over the years they performed numerous measurements of the movements of the Moon and the planets. The most industrious of all was the Dane Tycho Brahe,* who organized an industrial search for astronomical facts sponsored by his king. His measurements were the basis for the research of his young assistant, the Swabian astronomer Johannes Kepler** who found the first precise description of planetary Page 628 motion. In 1684, all observations of planets and stones were condensed into an astonishingly simple result by the English physicist Robert Hooke:*** every body of mass M attracts any other body towards its centre with an acceleration whose magnitude a is given by

$$a = G\frac{M}{r^2} \tag{30}$$

where r is the centre-to-centre distance of the two bodies. This is called the *universal 'law' of gravitation*, or universal gravity, because it is valid in general. The proportionality constant G is called the *gravitational constant*; it is one of the fundamental constants of nature, like the speed of light or the quantum of action. More about it will be said shortly. The effect of gravity thus decreases with increasing distance; gravity depends on the inverse square distance of the bodies under consideration. If bodies are small compared to the distance r, or if they are spherical, expression (30) is correct as it stands; for non-spherical shapes the acceleration has to be calculated separately for each part of the bodies and then added together.

This inverse square dependence is often called Newton's 'law' of gravitation, because the English physicist Isaac Newton proved more elegantly than Hooke that it agreed with all astronomical and terrestrial observations. Above all, however, he organized a better Ref. 89 public relations campaign, in which he falsely claimed to be the originator of the idea.

Newton published a simple proof showing that this description of astronomical motion also gives the correct description for stones thrown through the air, down here on 'father Earth'. To achieve this, he compared the acceleration a_{m} of the Moon with that of stones g. For the ratio between these two accelerations, the inverse square relation predicts a value $a_{\mathrm{m}}/g = R^2/d_{\mathrm{m}}^2$, where R is the radius of the Earth and d_{m} the distance of the Moon. The Moon's distance can be measured by triangulation, comparing the position of the Moon against the starry background from two different points on Earth.* The result

* Tycho Brahe (1546–1601), famous Danish astronomer, builder of Uraniaborg, the astronomical castle. He consumed almost 10 % of the Danish gross national product for his research, which produced the first star catalogue and the first precise position measurements of planets.

** Johannes Kepler (1571 Weil der Stadt–1630 Regensburg); after helping his mother defend herself in a trial where she was accused of witchcraft, he studied Protestant theology and became a teacher of mathematics, astronomy and rhetoric. His first book on astronomy made him famous, and he became assistant of Tycho Brahe and then, at his teacher's death, the Imperial Mathematician. He was the first to use mathematics in the description of astronomical observations, and introduced the concept and field of 'celestial physics'.

*** Robert Hooke, (1635–1703), important English physicist and secretary of the Royal Society. He also wrote the *Micrographia*, a beautifully illustrated exploration of the world of the very small.

* The first precise – but not the first – measurement was achieved in 1752 by the French astronomers Lalande and La Caille, who simultaneously measured the position of the Moon seen from Berlin and from Le Cap.

is $d_{\mathrm{m}}/R = 60 \pm 3$, depending on the orbital position of the Moon, so that an average ratio $a_{\mathrm{m}}/g = 3.6 \cdot 10^3$ is predicted from universal gravity. But both accelerations can also be measured directly. At the surface of the Earth, stones are subject to an acceleration due to gravitation with magnitude $g = 9.8\,\mathrm{m/s^2}$, as determined by measuring the time stones need to fall a given distance. For the Moon, the definition of acceleration, $a = \mathrm{d}v/\mathrm{d}t$, in the case of circular motion – roughly correct here – gives $a_{\mathrm{m}} = d_{\mathrm{m}}(2\pi/T)^2$, where $T = 2.4\,\mathrm{Ms}$ is the time the Moon takes for one orbit around the Earth.* The measurement of the radius of the Earth** yields $R = 6.4\,\mathrm{Mm}$, so that the average Earth–Moon distance is $d_{\mathrm{m}} = 0.38\,\mathrm{Gm}$. One thus has $a_{\mathrm{m}}/g = 3.6 \cdot 10^3$, in agreement with the above prediction. With this famous 'Moon calculation' we have thus shown that the inverse square property of gravitation indeed describes both the motion of the Moon and that of stones. You might want to deduce the value of GM. Challenge 218 n

From the observation that on the Earth all motion eventually comes to rest, whereas in the sky all motion is eternal, Aristotle and many others had concluded that motion in the sublunar world has different properties from motion in the translunar world. Several thinkers had criticized this distinction, notably the French philosopher and rector of the University of Paris, Jean Buridan.* The Moon calculation was the most important result showing this distinction to be wrong. This is the reason for calling the expression (30) the *universal* 'law' of gravitation. Ref. 92

This result allows us to answer another old question. Why does the Moon not fall from the sky? Well, the preceding discussion showed that *fall* is motion due to gravitation. Therefore the Moon actually *is* falling, with the peculiarity that instead of falling *towards* the Earth, it is continuously falling *around* it. Figure 63 presents the idea. The Moon is continuously missing the Earth.**

---

* This is deduced easily by noting that for an object in circular motion, the magnitude v of the velocity $\mathbf{v} = \mathrm{d}\mathbf{x}/\mathrm{d}t$ is given as $v = 2\pi r/T$. The drawing of the vector $\mathbf{v}$ over time, the so-called *hodograph*, shows that it behaves exactly like the position of the object. Therefore the magnitude a of the acceleration $\mathbf{a} = \mathrm{d}\mathbf{v}/\mathrm{d}t$ is given by the corresponding expression, namely $a = 2\pi v/T$. Challenge 215 n

** This is the hardest quantity to measure oneself. The most surprising way to determine the Earth's size is the following: watch a sunset in the garden of a house, with a stopwatch in hand. When the last ray of the Sun disappears, start the stopwatch and run upstairs. There, the Sun is still visible; stop the stopwatch when the Sun disappears again and note the time t. Measure the height distance h of the two eye positions where the Sun was observed. The Earth's radius R is then given by $R = k\,h/t^2$, with $k = 378 \cdot 10^6\,\mathrm{s}^2$. Ref. 90 Challenge 216 ny

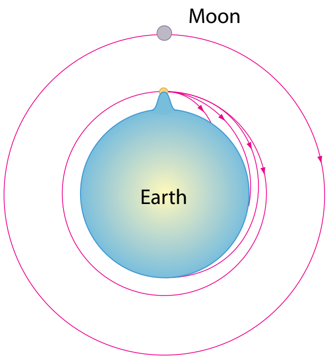
There is also a simple way to measure the distance to the Moon, once the size of the Earth is known. Take a photograph of the Moon when it is high in the sky, and call θ its zenith angle, i.e. its angle from the vertical. Make another photograph of the Moon a few hours later, when it is just above the horizon. On this picture, contrary to a common optical illusion, the Moon is smaller, because it is further away. With a drawing the reason for this becomes clear immediately. If q is the ratio of the two angular diameters, the Earth–Moon distance d_{m} is given by the relation $d_{\mathrm{m}}^2 = R^2 + [2Rq\cos\theta/(1-q^2)]^2$. Enjoy its derivation from the drawing. Ref. 91 Page 58 Challenge 217 ny

Another possibility is to determine the size of the Moon by comparing it to the size of the shadow of the Earth during an eclipse. The distance to the Moon is then computed from its angular size, about 0.5°.

* Jean Buridan (*c.* 1295 to *c.* 1366) was also one of the first modern thinkers to speculate on a rotation of the Earth about an axis.

** Another way to put it is to use the answer of the Dutch physicist Christiaan Huygens (1629–1695): the Moon does not fall from the sky because of the centrifugal acceleration. As explained on page 103, this explanation is nowadays out of favour at most universities.

There is a beautiful problem connected to the left part of the figure: Which points on the surface of the Earth can be reached by shooting from a mountain? And which points can be reached by shooting only horizontally? Ref. 93 Challenge 219 d



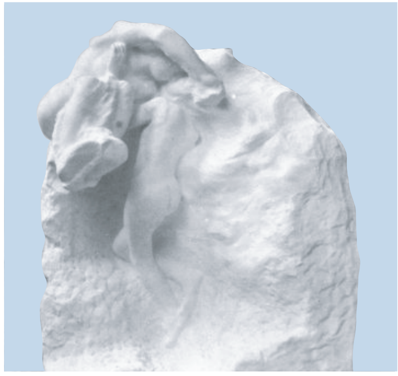


Figure 63 A physicist's and an artist's view of the fall of the Moon: a diagram by Christiaan Huygens (not to scale) and a marble by Auguste Rodin

Universal gravity also explains why the Earth and most planets are (almost) spherical. Since gravity increases with decreasing distance, a liquid body in space will always try to form a spherical shape. Seen on a large scale, the Earth is indeed liquid. We also know that the Earth is cooling down – that is how the crust and the continents formed. The sphericity of smaller solid objects encountered in space, such as the Moon, thus means that they used to be liquid in older times.

Properties of gravitation

Gravitation implies that the path of a stone is not a parabola, as stated earlier, but actually an ellipse around the centre of the Earth. This happens for exactly the same reason that the planets move in ellipses around the Sun. Are you able to confirm this statement? Challenge 220 ny

Universal gravitation allows us to solve a mystery. The puzzling acceleration value $g = 9.8\,\mathrm{m/s^2}$ we encountered in equation (5) is thus due to the relation

$$g = GM_{\mathrm{Earth}}/R^2_{\mathrm{Earth}} \,. \tag{31}$$

The equation can be deduced from equation (30) by taking the Earth to be spherical. The everyday acceleration of gravity g thus results from the size of the Earth, its mass, and the universal constant of gravitation G. Obviously, the value for g is almost constant on the surface of the Earth because the Earth is almost a sphere. Expression (31) also explains why g is smaller if one rises in the air, and the deviations of the shape of the Earth from sphericity explain why g is different at the poles and larger on a plateau. (What would one get on the Moon? On Mars? On Jupiter?) Challenge 221 ny

By the way, it is possible to devise a simple machine, other than a yo-yo, which slows down the effective acceleration of gravity by a known amount, so that one can measure its value more easily. Can you imagine it? Challenge 222 n

Note that 9.8 is roughly π^2. This is *not* a coincidence: the metre has been chosen in such a way to make this correct. The period T of a swinging pendulum, i.e. a back and forward swing, is given by* Challenge 223 ny

$$T = 2\pi\sqrt{\frac{l}{g}} \;, \tag{32}$$

where l is the length of the pendulum, and g is the gravitational acceleration. (The pendulum is assumed to be made of a mass attached to a string of negligible mass.) The oscillation time of a pendulum depends only on the length of the string and the planet it is located. If the metre had been defined such that $T/2 = 1\,\mathrm{s}$, the value of the normal acceleration g would have been exactly $\pi^2\,\mathrm{m/s^2}$. This was the first proposal for the definition of the metre; it was made in 1673 by Huygens and repeated in 1790 by Talleyrand, but was rejected by the conference that defined the metre because variations in the value of g with geographical position and temperature-induced variations of the length of a pendulum induce errors which are too large to yield a definition of useful precision.

Challenge 225 e

Finally, the proposal was made to define the metre as 1/40 000 000 of the circumference of the Earth through the poles, a so-called *meridian*. This proposal was almost identical to – but much more precise than – the pendulum proposal. The meridian definition of the metre was then adopted by the French national assembly on 26 March 1791, with the statement that 'a meridian passes under the feet of every human being, and all meridians are equal.' (Nevertheless, the distance from Equator to the poles is not exactly 10 Mm; that is a strange story. One of the two geographers who determined the size of the first metre stick was dishonest. The data he gave for his measurements – the general method of which is shown in Figure 64 – was invented. Thus the first official metre stick in Paris was shorter than it should be.)

Ref. 94

But we can still ask: Why does the Earth have the mass and size it has? And why does G have the value it has? The first question asks for a history of the solar system; it is still unanswered and a topic of research. The second question is addressed in Appendix B.

If all objects attract each other, that should also be the case for objects in everyday life. Gravity must also work *sideways*. This is indeed the case, even though the effects are so small that they were measured only long after universal gravity had predicted them. Measuring this effect allows the gravitational constant G to be determined.

Note that measuring the gravitational constant G is also the only way to determine the mass of the *Earth*. The first to do so, in 1798, was the English physicist Henry Cavendish; he used the machine, ideas and method of John Michell who died when attempting the experiment. Michell and Cavendish called the aim and result of his experiments 'weighing the Earth'. Are you able to imagine how they did it? The value found in experiments is

Challenge 226 n

$$G = 6.7 \cdot 10^{-11}\,\mathrm{Nm^2/kg^2} = 6.7 \cdot 10^{-11}\,\mathrm{m^3/kg\,s^2} \;. \tag{33}$$

* Formula (32) is noteworthy mainly for all that is missing. The period of a pendulum does *not* depend on the mass of the swinging body. In addition, the period of a pendulum does *not* depend on the amplitude. (This is true as long as the oscillation angle is smaller than about 15°.) Galileo discovered this as a student, when observing a chandelier hanging on a long rope in the dome of Pisa. Using his heartbeat as a clock he found that even though the amplitude of the swing got smaller and smaller, the time for the swing stayed the same.

A leg also moves like a pendulum, when one walks normally. Why then do taller people tend to walk faster?

Challenge 224 n

Cavendish's experiments were thus the first to confirm that gravity works also sideways.*

For example, two average people at a distance of 1 m apart feel an acceleration towards each other that is smaller than that exerted by a common fly when landing on the skin. Therefore we usually do not notice the attraction to other people. When we notice it, it is much stronger than that. This simple calculation thus proves that gravitation cannot be at the origin of people falling in love, and that sexual attraction is not of gravitational, but of different origin. This other interaction will be studied later in our walk; it is called *electromagnetism*.

Challenge 227 n

Figure 64 The measurements leading to the definition of the metre (© Ken Alder)

But gravity has more interesting properties to offer. The effects of gravitation can also be described by another observable, namely the *(gravitational) potential* φ. We then have the simple relation that the acceleration is given by the *gradient* of the potential

$$\mathbf{a} = -\nabla\varphi \qquad \text{or} \qquad \mathbf{a} = -\operatorname{grad}\varphi\ . \tag{34}$$

The gradient is just a learned term for 'slope along the steepest direction'. It is defined for any point on a slope, is large for a steep one and small for a shallow one and it points in the direction of steepest ascent, as shown in Figure 65. The gradient is abbreviated ∇, pronounced 'nabla' and is mathematically defined as the vector $\nabla\varphi = (\partial\varphi/\partial x, \partial\varphi/\partial y, \partial\varphi/\partial z) = \operatorname{grad}\varphi$. The minus sign in the above definitions is introduced by convention, in order to have higher potential values at larger heights.* For a point-like or a spherical body of mass M, the potential φ is

$$\varphi = -G\frac{M}{r}\ . \tag{35}$$

A potential considerably simplifies the description of motion, since a potential is additive: given the potential of a point particle, one can calculate the potential and then the motion around any other, irregularly shaped object.**

* Henry Cavendish (1731–1810) was one of the great geniuses of physics; rich and solitary, he found many rules of nature, but never published them. Had he done so, his name would be much more well-known.

John Michell (1724–1793) was church minister, geologist and amateur astronomer.

* In two or more dimensions slopes are written $\partial\varphi/\partial z$ – where ∂ is still pronounced 'd' – because in those cases the expression $d\varphi/dz$ has a slightly different meaning. The details lie outside the scope of this walk.

** Alternatively, for a general, extended body, the potential is found by requiring that the *divergence* of its gradient is given by the mass (or charge) density times some proportionality constant. More precisely, one has

$$\Delta\varphi = 4\pi G\rho \tag{36}$$

where $\rho = \rho(\mathbf{x}, t)$ is the mass volume density of the body and the operator Δ, pronounced 'delta', is defined as $\Delta f = \nabla\nabla f = \partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 + \partial^2 f/\partial z^2$. Equation (36) is called the *Poisson equation* for the potential φ. It is named after Siméon-Denis Poisson (1781–1840), eminent French mathematician and physicist. The

The potential φ is an interesting quantity; with a single number at every position in space we can describe the vector aspects of gravitational acceleration. It automatically describes that gravity in New Zealand acts in the opposite direction to gravity in Paris. In addition, the potential suggests the introduction of the so-called *potential energy* U by setting

$$U = m\varphi \qquad (37)$$

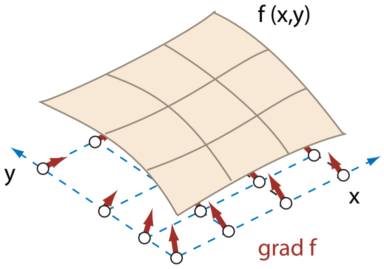


Figure 65 The potential and the gradient

and thus allowing us to determine the change of *kinetic* energy T of a body falling from a point 1 to a point 2 via

$$T_1 - T_2 = U_2 - U_1 \quad \text{or} \quad \frac{1}{2}m_1\mathbf{v}_1{}^2 - \frac{1}{2}m_2\mathbf{v}_2{}^2 = m\varphi_2 - m\varphi_1 \,. \qquad (38)$$

In other words, the total energy, defined as the sum of kinetic and potential energy, is *conserved* in motion due to gravity. This is a characteristic property of gravitation. Not all accelerations can be derived from a potential; systems with this property are called *conservative*. The accelerations due to friction are not conservative, but those due to electromagnetism are.

Interestingly, the number of dimensions of space d is coded into the potential of a spherical mass: its dependence on the radius r is in fact $1/r^{d-2}$. The exponent $d-2$ has been checked experimentally to high precision; no deviation of d from 3 has ever been found.

Challenge 229 ny

Ref. 96

The concept of potential helps in understanding the *shape* of the Earth. Since most of the Earth is still liquid when seen on a large scale, its surface is always horizontal with respect to the direction determined by the combination of the accelerations of gravity and rotation. In short, the Earth is *not* a sphere. It is not an ellipsoid either. The mathematical shape defined by the equilibrium requirement is called a *geoid*. The geoid shape differs from a suitably chosen ellipsoid by at most 50 m. Can you describe the geoid mathematically? The geoid is an excellent approximation to the actual shape of the Earth; sea level differs from it by less than 20 metres. The differences can be measured with satellite radar and are of great interest to geologists and geographers. For example, it turns out that the South Pole is nearer to the equatorial plane than the North Pole by about 30 m. This is probably due to the large land masses in the northern hemisphere.

Ref. 97

Ref. 98

Challenge 230 ny

positions at which ρ is not zero are called the *sources* of the potential. The so-called source term $\Delta\varphi$ of a function is a measure for how much the function $\varphi(x)$ at a point x differs from the average value in a region around that point. (Can you show this, by showing that $\Delta\varphi \approx \bar{\varphi} - \varphi(x)$?) In other words, the Poisson equation (36) implies that the actual value of the potential at a point is the same as the average value around that point minus the mass density multiplied by $4\pi G$. In particular, in the case of empty space the potential at a point is equal to the average of the potential around that point.

Challenge 228 ny

Often the concept of *gravitational field* is introduced, defined as $\mathbf{g} = -\nabla\varphi$. We avoid this in our walk, because we will discover that following the theory of relativity gravity is not due to a field at all; in fact even the concept of gravitational potential turns out to be only an approximation.

Above we saw how the inertia of matter, Page 90 through the so-called 'centrifugal force', increases the radius of the Earth at the Equator. In other words, the Earth is *flattened* at the poles. The Equator has a radius a of 6.38 Mm, whereas the distance b from the poles to the centre of the Earth is 6.36 Mm. The precise flattening $(a - b)/a$ Appendix B has the value $1/298.3 = 0.0034$. As a result, the top of Mount Chimborazo in Ecuador, even though its height is only 6267 m above sea level, is about 20 km farther away from the centre of the Earth than the top of Mount Sagarmatha* in Nepal, whose height above sea level is 8850 m. The top of Mount Chimborazo is in fact the surface point most distant from the centre of the Earth.

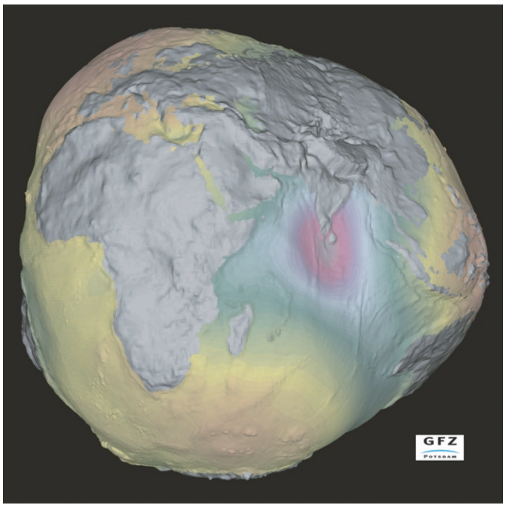


Figure 66 The shape of the Earth, with exaggerated height scale (© GeoForschungsZentrum Potsdam)

As a consequence, if the Earth stopped rotating (but kept its shape), the water of the oceans would flow north; all of Europe would be under water, except for the few mountains of the Alps higher than about 4 km. The northern parts of Europe would be covered by between 6 km and 10 km of water. Mount Sagarmatha would be over 11 km above sea level. If one takes into account the resulting shape change of the Earth, the numbers come out smaller. In addition, the shape change would produce extremely strong earthquakes and storms. As long as all these effects are lacking, we are *sure* that the Sun will indeed Page 83 rise tomorrow, despite what some philosophers might pretend.

Dynamics: how do things move in various dimensions?

Let us give a short summary. If a body can only move along a (possibly curved) line, the concepts of kinetic and potential energy are sufficient to determine the way it moves. In short, motion in *one dimension* follows directly from energy conservation.

If more *two spatial dimensions* are involved, energy conservation is not sufficient to determine how a body moves. If a body can move in *two dimensions*, *and* if the forces involved are internal, the conservation of angular momentum can be used. The full motion in two dimensions follows from energy and angular momentum conservation. For example, the properties of free fall follow from energy and angular momentum conservation. (Are you able to show this?) Challenge 231 ny

In the case of motion in *three dimensions*, the general rule for determining motion is necessary. It turns out that all motion follows from a simple principle: the time average of the difference between kinetic and potential energy must be as small as possible. This is called the *least action principle*. We will explain the details of this calculation method later on. Page 161

* Mount Sagarmatha is sometimes also called Mount Everest.

For simple gravitational motions, motion is two-dimensional. Most three-dimensional problems are outside the scope of this text; in fact, some of these problems are still subjects of research. In this adventure, we will explore three-dimensional motion only for selected cases that provide important insights.

Gravitation in the sky

The expression for the acceleration due to gravity $a = GM/r^2$ also describes the motion of all the planets around the Sun. Anyone can check that the planets always stay within the *zodiac*, a narrow stripe across the sky. The centre line of the zodiac gives the path of the Sun and is called the *ecliptic*, since the Moon must be located on it to produce an eclipse. But the detailed motion of the planets is not easy to describe.* A few generations before Hooke, the Swabian astronomer Johannes Kepler had deduced several 'laws' in his painstaking research about the movements of the planets in the zodiac. The three main ones are:

- Planets move on ellipses with the Sun located at one focus (1609);
- Planets sweep out equal areas in equal times (1609);
- All planets have the same ratio T^2/d^3 between the orbit duration T and the semimajor axis d (1619).

The main results are given in Figure 67. The sheer work required to deduce the three 'laws' was enormous. Kepler had no calculation machine available, not even a slide rule. The calculation technology he used was the recently discovered logarithms. Anyone who has used tables of logarithms to actually perform calculations can get a feeling for the amount of work behind these three discoveries.

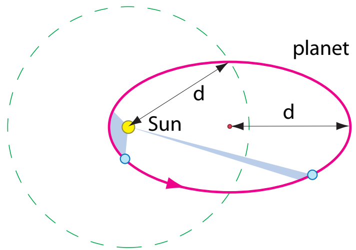


Figure 67 The motion of a planet around the Sun, showing its semimajor axis d, which is also the spatial average of its distance from the Sun

The law about equal swept areas implies that planets move faster when they are near the Sun. It is a way to state the conservation of angular momentum. But now comes the central point. All the huge work by Brahe and Kepler can be summarized in the expression $a = GM/r^2$. Can you confirm that all three laws follow from Hooke's expression of universal gravity? Publishing this result was the main achievement of Newton. Try to repeat his achievement; it will show you the difficulties, but also the possibilities of physics, and the joy that puzzles offer.

Challenge 232 ny

Newton solved the puzzle with geometric drawing. Newton was not able to write down, let alone handle, differential equations at the time he published his results on gravitation. In fact, it is well known that Newton's notation and calculation methods were poor. (Much poorer than yours!) The English mathematician Godfrey Hardy** used to say that the

Ref. 21

* The apparent height of the ecliptic changes with the time of the year and is the reason for the changing seasons. Therefore seasons are a gravitational effect as well.

** Godfrey Harold Hardy (1877–1947) was an important English number theorist, and the author of the well-known *A Mathematician's Apology*. He also 'discovered' the famous Indian mathematician Srinivasa Ramanujan, bringing him to Britain.

insistence on using Newton's integral and differential notation, rather than the earlier and better method, still common today, due to his rival Leibniz – threw back English mathematics by 100 years.

Kepler, Hooke and Newton became famous because they brought order to the description of planetary motion. This achievement, though of small practical significance, was widely publicized because of the age-old prejudices linked to astrology.

However, there is more to gravitation. Universal gravity explains the motion and shape of the Milky Way and of the other galaxies, the motion of many weather phenomena and explains why the Earth has an atmosphere but the Moon does not. (Can you do the same?) In fact, universal gravity explains much more about the Moon.

Challenge 233 ny

The Moon

How long is a day on the Moon? The answer is roughly 14 Earth-days. That is the time that it takes for the Moon to see the Sun again in the same position.

One often hears that the Moon always shows the same side to the Earth. But this is wrong. As one can check with the naked eye, a given feature in the centre of the face of the Moon at full Moon is not at the centre one week later. The various motions leading to this change are called *librations*; they appear mainly because the Moon does not describe a circular, but an elliptical orbit around the Earth and because the axis of the Moon is slightly inclined compared to that of its rotation around the Earth. As a result, only around 45 % of the Moon's surface is permanently hidden from Earth.

The first photographs of the hidden areas were taken in the 1960s by a Soviet artificial satellite. The surface is much more irregular than the visible one, as the hidden side is the one which intercepts most asteroids attracted by the Earth. Thus the gravitation of the Moon helps to deflect asteroids from the Earth. The number of animal life extinctions is thus reduced to a small, but not negligible number. In other words, the gravitational attraction of the Moon has saved the human race from extinction many times over.*

The trips to the Moon in the 1970s also showed that the Moon originated from the Earth itself: long ago, an object hit the Earth almost tangentially and threw a sizeable fraction of material up into the sky. This is the only mechanism able to explain the large size of the Moon, its low iron content, as well as its general material composition.

Ref. 99

The Moon is receding from the Earth at 3.8 cm a year. This result confirms the old deduction that the tides slow down the Earth's rotation. Can you imagine how this measurement was performed?** Since the Moon slows down the Earth, the Earth also changes shape due to this effect. (Remember that the shape of the Earth depends on its rotation speed.) These changes in shape influence the tectonic activity of the Earth, and maybe also the drift of the continents.

Challenge 234 n

The Moon has many effects on animal life. A famous example is the midge Clunio, which lives on sea coasts with pronounced tides. Clunio lives between six and twelve

Ref. 100

* The web pages http://cfa-www.harvard.edu/iau/lists/Closest.html and InnerPlot.html give an impression of the number of objects which almost hit the Earth every year. Without the Moon, we would have many additional catastrophes.

** If you want to read about the motion of the Moon in all its fascinating details, have a look at Martin C. Gutzwiller, Moon-Earth-sun: the oldest three body problem, *Reviews of Modern Physics* **70**, pp. 589–639, 1998.

weeks as a larva then hatches and lives only one or two hours as adult flying insect, during which time it reproduces. The reproduction is only successful if the midge hatches during the low tide phase of a spring tide. Spring tides are the especially strong tides during the full and new moons, when the solar and lunar effects add, and occur only every 14.8 days. In 1995, Dietrich Neumann showed that the larvae have two built-in clocks, a circadian and a circalunar one, which together control the hatching to precisely those few hours when the insect can reproduce. He also showed that the circalunar clock is synchronized by the brightness of the Moon at night. In other words, the larvae watch the Moon at night and then decide when to hatch: they are the smallest known astronomers.

If insects can have circalunar cycles, it should come as no surprise that women also have such a cycle. However, in this case the origin of the cycle length is still unknown. Ref. 101

The Moon also helps to stabilize the tilt of the Earth's axis, keeping it more or less fixed relative to the plane of motion around the Sun. Without the Moon, the axis would change its direction irregularly, we would not have a regular day and night rhythm, we would have extremely large climate changes, and the evolution of life would have been impossible. Without the Moon, the Earth would also rotate much faster and we would have much less friendly weather. The Moon's main remaining effect on the Earth, the precession of its axis, is responsible for the ice ages. Ref. 102 Ref. 103 Page 94

Furthermore, the Moon shields the Earth from cosmic radiation by greatly increasing the Earth's magnetic field. In other words, the Moon is of central importance for the evolution of life. Understanding how often Earth-sized planets have Moon-sized satellites is thus important for the estimation of the probability that life exists on other planets. So far, it seems that large satellites are rare; there are only four known moons that are larger than that of the Earth, but they circle much larger planets, namely Jupiter and Saturn. Indeed, the formation of satellites is still an area of research. But let us return to the effects of gravitation in the sky. Ref. 104

Orbits

The path of one body orbiting another under the influence of gravity is an ellipse with the central body at one focus. A circular orbit is also possible, a circle being a special case of an ellipse. Single encounters of two objects can also be parabolas or hyperbolas, as shown in Figure 68. Circles, ellipses, parabolas and hyperbolas are collectively known as *conic sections*. Indeed each of these curves can be produced by cutting a cone with a knife. Are you able to confirm this? Challenge 235 e

If orbits are mostly ellipses, it follows that comets return. The English astronomer Edmund Halley (1656–1742) was the first to draw this conclusion and to predict the return of a comet. It arrived at the predicted date in 1756, and is now named after him. The period of Halley's comet is between 74 and 80 years; the first recorded sighting was 22 centuries ago, and it has been seen at every one of its 30 passages since, the last time in 1986.

Depending on the initial energy and the initial angular momentum of the body with respect to the central planet, there are two additional possibilities: *parabolic* paths and *hyperbolic* paths. Can you determine the conditions on the energy and the angular momentum for these paths to appear? Challenge 236 ny

In practice, parabolic paths do not exist in nature. (Though some comets seem to approach this case when moving around the Sun; almost all comets follow elliptical paths).

Hyperbolic paths do exist; artificial satellites follow them when they are shot towards a planet, usually with the aim of changing the direction of their journey across the solar system.

Why does the inverse square law lead to conic sections? First of all, for two bodies, the total angular momentum L is a constant:

$$L = mr^2\dot{\varphi} \qquad (39)$$

and therefore the motion lies in a plane. Also the energy E is a constant

$$E = \frac{1}{2}m\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + \frac{1}{2}m\left(r\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^2 - G\frac{mM}{r} \ . \qquad (40)$$

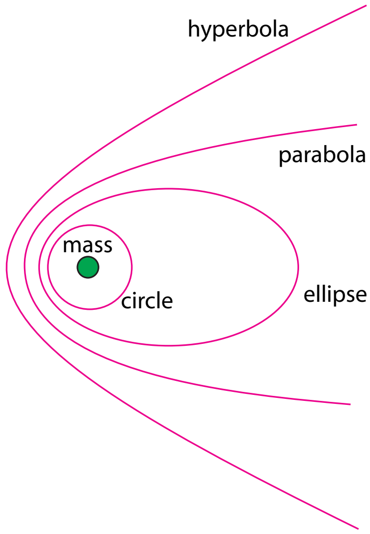


Figure 68 The possible orbits due to universal gravity

Challenge 237 ny Together, the two equations imply that

$$r = \frac{L^2}{Gm^2M}\,\frac{1}{1+\sqrt{1+\frac{2EL^2}{G^2m^3M^2}}\cos\varphi} \ . \qquad (41)$$

Now, any curve defined by the general expression

$$r = \frac{C}{1+e\cos\varphi} \quad \text{or} \quad r = \frac{C}{1-e\cos\varphi} \qquad (42)$$

is an ellipse for $0 < e < 1$, a parabola for $e = 1$ and a hyperbola for $e > 1$, one focus being at the origin. The quantity e, called the *eccentricity*, describes how squeezed the curve is. In other words, a body in orbit around a central mass follows a conic section.

If more than two objects move under mutual gravitation, many additional possibilities for motions appear. The classification and the motions are quite complex. In fact, this so-called *many-body problem* is still a topic of research, and the results are mathematically fascinating. Let us look at a few examples.

When several planets circle a star, they also attract each other. Planets thus do not move in perfect ellipses. The largest deviation is a perihelion shift, as shown in Figure 48. It is observed for Mercury and a few other planets, including the Earth. Other deviations from elliptical paths appear during a single orbit. In 1846, the observed deviations of the motion of the planet Uranus from the path predicted by universal gravity were used to predict the existence of another planet, Neptune, which was discovered shortly afterwards.

Page 76 We have seen that mass is always positive and that gravitation is thus always attractive; there is no *antigravity*. Can gravity be used for *levitation* nevertheless, maybe using more than two bodies? Yes; there are two examples.* The first are the geostationary satellites, which are used for easy transmission of television and other signals from and towards Earth.

* Levitation is discussed in detail on page 555.

The *Lagrangian libration points* are the second example. Named after their discoverer, these are points in space near a two-body system, such as Moon–Earth or Earth–Sun, in which small objects have a stable equilibrium position. A general overview is given in Figure 69. Can you find their precise position, not forgetting to take rotation into account? There are three additional Lagrangian points on the Earth–Moon line. How many of them are stable?

Challenge 238 ny

Challenge 239 ny

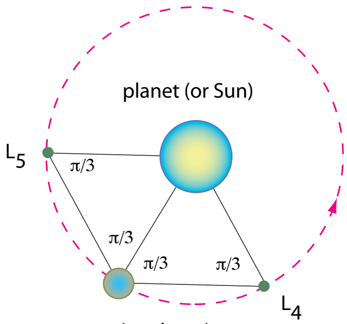


Figure 69 The two stable Lagrangian points

There are thousands of asteroids, called *Trojan asteroids*, at and around the Lagrangian points of the Sun–Jupiter system. In 1990, a Trojan asteroid for the Mars–Sun system was discovered. Finally, in 1997, a Trojan asteroid was found which follows the Earth in its way around the Sun. This second companion of the Earth has a diameter of 5 km. Similarly, on the main Lagrangian points of the Earth–Moon system a high concentration of dust has been observed.

Ref. 105

To sum up, the single equation $\mathbf{a} = -GM\mathbf{r}/r^3$ correctly describes a large number of phenomena in the sky. The first person to make clear that the expression describes *everything* happening in the sky was Pierre Simon Laplace* in his famous treatise *Traité de mécanique céleste*. When Napoleon told him that he found no mention about the creator in the book, Laplace gave a famous, one sentence summary of his book: 'I did not need this hypothesis any more'. In particular, Laplace studied the stability of the solar system, the eccentricity of the lunar orbit, the eccentricities of the planetary orbits, always getting full equivalence between calculation and measurement.

These results are quite a feat for the simple expression of universal gravitation; they also explain why it is called 'universal'. But how precise is the formula? Since astronomy allows the most precise measurements of gravitational motion, it also provides the most stringent tests. In 1882, Simon Newcomb (1835–1909) repeated Laplace's analysis and concluded after intensive study that there was only one known example of discrepancy from universal gravity, namely one observation for the planet Mercury. (Nowadays a few more are known.) The point of smallest distance to the Sun of the orbit of planet Mercury, its *perihelion*, changes with a rate slightly smaller than the predicted one: the tiny difference is around 43″ per century. The study of motion had to wait for Albert Einstein to explain it.

Ref. 106

Tides

Why do physics texts always talk about tides? Because, as general relativity will show, tides prove that space is curved! It is thus useful to study them a bit in more detail. Gravitation describes the sea tides as results of the attraction of the ocean water by the Moon and the Sun. Tides are interesting; even though the amplitude of the tides is only about 0.5 m on the open sea, it can be up to 20 m at special places near the coast. Can you imagine why? The *soil* is also lifted and lowered by the Sun and the Moon, by about 0.3 m, as satel-

Challenge 240 n

* Pierre Simon Laplace (b. 1749 Beaumont-en-Auge, d. 1827 Paris), important French mathematician. His treatise appeared in 5 volumes between 1798 and 1825. He was the first to propose that the solar system was formed from a rotating gas cloud, and one of the first people to imagine and explore black holes.

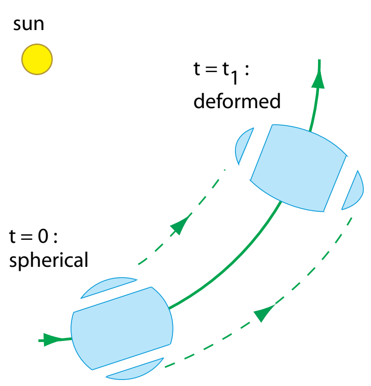


Figure 70 Tidal deformations due to gravity

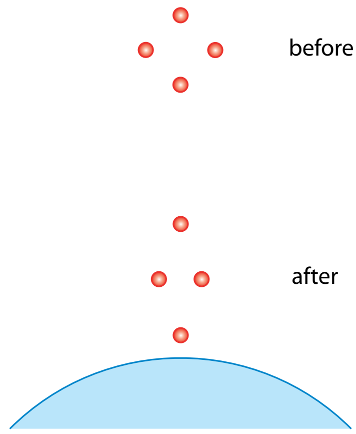


Figure 71 The origin of tides

Ref. 36 lite measurements show. Even the *atmosphere* is subject to tides, and the corresponding
Ref. 107 pressure variations can be filtered out from the weather pressure measurements.

Tides appear for any *extended* body moving in the gravitational field of another. To understand the origin of tides, picture a body in orbit, like the Earth, and imagine its components, such as the segments of Figure 70, as being kept together by springs. Universal gravity implies that orbits are slower the more distant they are from a central body. As a result, the segment on the outside of the orbit would like to be slower than the central one; through the springs it is *pulled* by the rest of the body. In contrast, the inside segment would like to orbit more rapidly and is thus *retained* by the others. Being slowed down, the inside segments wants to fall towards the Sun. In sum, both segments feel a pull away from the centre of the body, until the springs stop the deformation. Therefore, *extended bodies are deformed in the direction of the field inhomogeneity.*

For example, as a result of tidal forces, the Moon always points with (roughly) the same face to the Earth. In addition, its radius towards the Earth is larger by about 5 m than the radius perpendicular to it. If the inner springs are too weak, the body is torn into pieces; in this way a *ring* of fragments can form, such as the asteroid ring between Mars and Jupiter or the rings around Saturn.

Let us return to the Earth. If a body is surrounded by water, it will form bulges in the direction of the applied gravitational field. In order to measure and compare the strength of the tides from the Sun and the Moon, we reduce tidal effects to their bare minimum, as shown in Figure 71. Tides appear because nearby points falling together approach or diverge, depending on their relative position. Tides thus depend on the change of acceleration with distance; in other words, this *relative* acceleration is proportional to the derivative of the gravitational acceleration.

Using the numbers from Appendix B, the gravitational accelerations from the Sun and

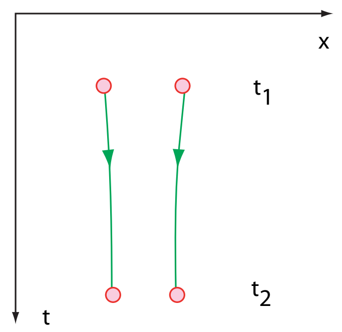


Figure 72 Particles falling side by side approach over time

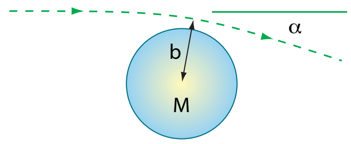


Figure 73 Masses bend light

the Moon measured on Earth are

$$a_{\text{Sun}} = \frac{GM_{\text{Sun}}}{d_{\text{Sun}}^2} = 5.9\,\text{mm/s}^2$$
$$a_{\text{Moon}} = \frac{GM_{\text{Moon}}}{d_{\text{Moon}}^2} = 0.033\,\text{mm/s}^2 \qquad (43)$$

and thus the attraction from the Moon is about 178 times weaker than that from the Sun.

When two nearby bodies fall near a large mass, the relative acceleration is proportional to their distance, and follows $da = da/dr\ dr$. The proportionality factor $da/dr = \nabla a$, called the *tidal acceleration* (gradient), is the true measure of tidal effects. Near a large spherical mass M, it is given by

Challenge 241 e

$$\frac{da}{dr} = -\frac{2GM}{r^3} \qquad (44)$$

which yields the values

$$\frac{da_{\text{Sun}}}{dr} = -\frac{2GM_{\text{Sun}}}{d_{\text{Sun}}^3} = -0.8 \cdot 10^{-13}\,/\text{s}^2$$
$$\frac{da_{\text{Moon}}}{dr} = -\frac{2GM_{\text{Moon}}}{d_{\text{Moon}}^3} = -1.7 \cdot 10^{-13}\,/\text{s}^2\ . \qquad (45)$$

In other words, despite the much weaker pull of the Moon, its tides are predicted to be over *twice as strong* as the tides from the Sun; this is indeed observed. When Sun, Moon and Earth are aligned, the two tides add up; these so-called *spring tides* are especially strong and happen every 14.8 days, at full and new moon.

Tides also produce *friction*. The friction leads to a slowdown of Earth's rotation. Nowadays, the slowdown can be measured by precise clocks (even though short time variations due to other effects, such as the weather, are often larger). The results fit well with fossil results showing that 400 million years ago, in the Devonian, a year had 400 days, and a day about 22 hours. It is also estimated that 900 million years ago, each of

Ref. 70

the 481 days of a year were 18.2 hours long. The friction at the basis of this slowdown also results in an increase of the distance of the Moon by about 3.8 cm per year. Are you able to explain why?
Challenge 242 n

In summary, tides are due to relative accelerations of nearby mass points. This has an important consequence. In the chapter on general relativity we will find that time multiplied by the speed of light plays the same role as length. Time then becomes an additional dimension, as shown in Figure 72. Using this similarity, two free particles moving in the same direction correspond to parallel lines in space-time. Two particles falling side-by-side also correspond to parallel lines. Tides show that such particles approach each other. In other words, tides imply that parallel lines approach each other. But parallel lines can approach each other *only* if space-time is curved. In short, tides imply *curved* space-time and space. This simple reasoning could have been performed in the eighteenth century; however, it took another 200 years and Albert Einstein's genius to uncover it.
Page 344
Page 383

Can light fall?

> Die Maxime, jederzeit selbst zu denken, ist die Aufklärung.
>
> Immanuel Kant*

Towards the end of the seventeenth century people discovered that light has a finite velocity – a story which we will tell in detail later on. An entity that moves with infinite velocity cannot be affected by gravity, as there is no time to produce an effect. An entity with a finite speed, however, should feel gravity and thus fall.
Page 249

Does the speed increase when light reaches the surface of the Earth? For almost three centuries people had no means means of detecting any such effect; so the question was not investigated. Then, in 1801, the Prussian astronomer Johann Soldner (1776–1833) was the first to put the question in a different way. Being an astronomer, he was used to measuring stars and their observation angles. He realized that light passing near a massive body would be *deflected* due to gravity.
Ref. 108

Soldner studied a body on a hyperbolic path, moving with velocity c past a spherical mass M at distance b (measured from the centre), as shown in Figure 73. Soldner deduced the deflection angle
Challenge 243 ny

$$\alpha_{\text{univ. grav.}} = \frac{2}{b}\frac{GM}{c^2} \; . \qquad (46)$$

One sees that the angle is largest when the motion is just grazing the mass M. For light deflected by the mass of the Sun, the angle turns out to be at most a tiny $0.88'' = 4.3\,\mu\text{rad}$. In Soldner's time, this angle was too small to be measured. Thus the issue was forgotten. Had it been pursued, general relativity would have started as an experimental science, and not as a theoretical effort by Albert Einstein! Why? The value just calculated is *different* from the measured value. The first measurement took place in 1919;** it found the correct dependence on the distance, but found a deflection up to $1.75''$, exactly the double of expression (46). The reason is not easy to find; in fact, it is due to the curvature of space, as we will see. In summary, light can fall, but the issue conceals some surprises.
Page 375

* The maxim to think at all times for oneself is the enlightenment.

** By the way, how would you measure the deflection of light near the bright Sun?
Challenge 244 ny

What is mass? – Again

Mass describes how an object interacts with others. In our walk, we have encountered two of its aspects. *Inertial mass* is the property that keeps objects moving and which offers resistance to change of their motion. *Gravitational mass* is the property responsible for the acceleration of bodies nearby (the active aspect) or of being accelerated by objects nearby (the passive aspect). For example, the active aspect of the mass of the Earth determines the surface acceleration of bodies; the passive aspect of the bodies allows us to weigh them in order to measure their mass using distances only, e.g. on a scale or a balance. The gravitational mass is the basis of *weight*, the difficulty of lifting things.*

Is the gravitational mass of a body equal to the inertial mass? A rough answer is given by the experience that an object that is difficult to move is also difficult to lift. The simplest experiment is to take two bodies of different mass and let them fall. If the acceleration is the same for all bodies, inertial mass is equal to (passive) gravitational mass, because in the relation $ma = \nabla(GMm/r)$ the left m is actually the inertial mass, and the right m is actually the gravitational mass.

But in the seventeenth century Galileo had made widely known an even older argument showing without a single experiment that the acceleration is indeed the same for all bodies. If larger masses fell more rapidly than smaller ones, then the following paradox would appear. Any body can be seen as composed from a large fragment attached to a small fragment. If small bodies really fell less rapidly, the small fragment would slow the large fragment down, so that the complete body would have to fall *less* rapidly than the larger fragment (or break into pieces). At the same time, the body being larger than its fragment, it should fall *more* rapidly than that fragment. This is obviously impossible: all masses must fall with the same acceleration.

Many accurate experiments have been performed since Galileo's original discussion. In all of them the independence of the acceleration of free fall from mass and material
Ref. 109
composition has been confirmed with the precision they allowed. In other words, as far as we can tell, the gravitational mass and the inertial mass are *identical*. What is the origin of this mysterious equality?

This so-called 'mystery' is a typical example of disinformation, now common across
Page 72
the whole world of physics education. Let us go back to the definition of mass as negative inverse acceleration ratio. We mentioned that the physical origins of the accelerations do not play a role in the definition because the origin does not appear in the expression. In other words, the value of the mass is by definition independent of the interaction. That means in particular that inertial mass, based on electromagnetic interaction, and gravitational mass are identical *by definition*.

We also note that we have never defined a separate concept of 'passive gravitational mass'. The mass being accelerated by gravitation is the inertial mass. Worse, there is no
Challenge 246 ny
way to define a 'passive gravitational mass' at all. Try it! All methods, such as weighing an object, cannot be distinguished from those that determine inertial mass from its reaction to acceleration. Indeed, all methods to measure mass use non-gravitational mechanisms. Scales are good examples.

If the 'passive gravitational mass' were different from inertial mass, we would have strange consequences. For those bodies for which it were different we would get into

Challenge 245 ny
* What are the values shown by a balance for a person of 85 kg juggling three balls of 0.3 kg each?

trouble with energy conservation. Also assuming that 'active gravitational mass' differs from inertial mass get us into trouble.

Another way to look at the issue is the following. How could 'gravitational mass' differ from inertial mass? Would the difference depend on relative velocity, time, position, composition or on mass itself? Each of these possibilities contradicts either energy or momentum conservation.

No wonder that all measurements confirm the equality of all mass types. The issue is
Page 361
usually resurrected in general relativity, with no new results. 'Both' masses remain equal; mass is a unique property of bodies. Another issue remains, though. What is the *origin* of mass? Why does it exist? This simple but deep question cannot be answered by classical physics. We will need some patience to find out.

Curiosities and fun challenges about gravitation

> Fallen ist weder gefährlich noch eine Schande; Liegen bleiben ist beides.*
>
> Konrad Adenauer

- The inverse square expression of universal gravity has a limitation: it does not allow to make sensible statements about the matter in the universe. Universal gravity does predict that a homogeneous mass distribution is unstable; indeed, an inhomogeneous distribution is observed. However, universal gravity does not predict the average mass density, the darkness at night, the observed speeds of the distant galaxies, etc. In fact, not a single property of the universe is predicted. To do this, we need general relativity.
- Imagine that you have 12 coins of identical appearance, of which one is a forgery. The forged one has a different mass from the 11 genuine ones. How can you decide which is the forged one and whether it is lighter or heavier, using a simple balance only three times?
Challenge 247 e
- For a physicist, antigravity is repulsive gravity; it does not exist in nature. Nevertheless, the term 'antigravity' is used incorrectly by many people, as a short search on the internet shows. These people call any effect that *overcomes* gravity in this way. However, this definition implies that tables and chairs are antigravity devices. Following the definition, most of the wood, steel and concrete producers are in the antigravity business. The internet definition makes absolutely no sense.

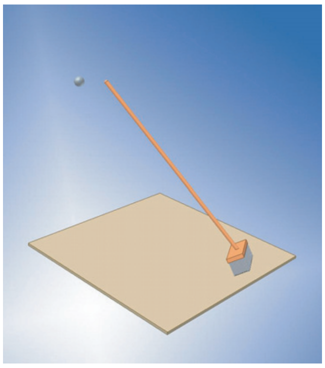


Figure 74 Brooms fall more rapidly than stones (© Luca Gastaldi)

- Do all objects on Earth fall with the same acceleration of $9.8\,\mathrm{m/s^2}$, assuming that air resistance can be neglected? No; every housekeeper knows that. You can check this by yourself. A broom angled at around 35° hits the floor earlier than a stone, as the impact noises tell. Are you able to explain why?
Challenge 248 n

* 'Falling is neither dangerous nor a shame; to keep lying is both.' Konrad Adenauer (b. 1876 Köln, d. 1967 Rhöndorf), German chancellor.

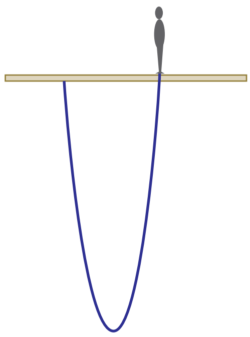


Figure 75 The start situation for bungee jumping

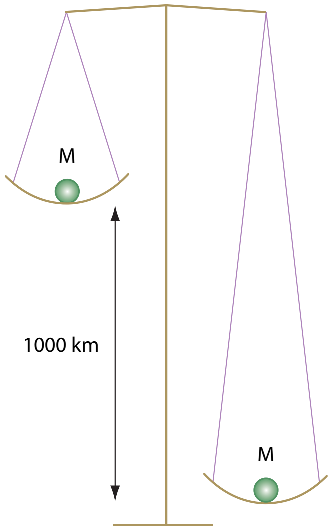


Figure 76 An honest balance?

▪ Also Bungee jumpers are accelerated more strongly than g. For a rubber of mass m and a jumper of mass M, the maximum acceleration a is

$$a = g\left(1 + \frac{1}{8}\frac{m}{M}\left(4 + \frac{m}{M}\right)\right) . \qquad (47)$$

Challenge 249 n Can you deduce the relation from Figure 75?

Challenge 250 n ▪ Guess: what is the weight of a ball of cork with a radius of 1 m?

Challenge 251 n ▪ Guess: A heap of 1000 iron balls with 1 mm of diameter is collected. What is its mass?

Challenge 252 n ▪ How can you use your observations made during travels to show that the Earth is not flat?

▪ Is the acceleration due to gravity constant? Not really. Every day, it is estimated that 10^8 kg fall onto the Earth as meteorites.

Challenge 253 n ▪ Both the Earth and the Moon attract bodies. The centre of mass of the Earth–Moon system is 4800 km away from the centre of the Earth, quite near its surface. Why do bodies on Earth still fall towards the centre of the Earth?

Challenge 254 ny ▪ Does every spherical body fall with the same acceleration? No. If the weight of the object is comparable to that of the Earth, the distance decreases in a different way. Can you confirm this statement? What then is wrong about Galileo's argument about the constancy of acceleration of free fall?

Challenge 255 n ▪ It is easy to lift a mass of a kilogram on a table. Twenty kilograms is tougher. A thousand is impossible. However, $6 \cdot 10^{24}$ kg is easy. Why?

Challenge 256 ny ▪ The strength ratio between the tides of Moon and Sun is roughly 7/3. Is it true that this is also the ratio between the mass densities of the two bodies?

▪ The friction between the Earth and the Moon slows down the rotation of both. The

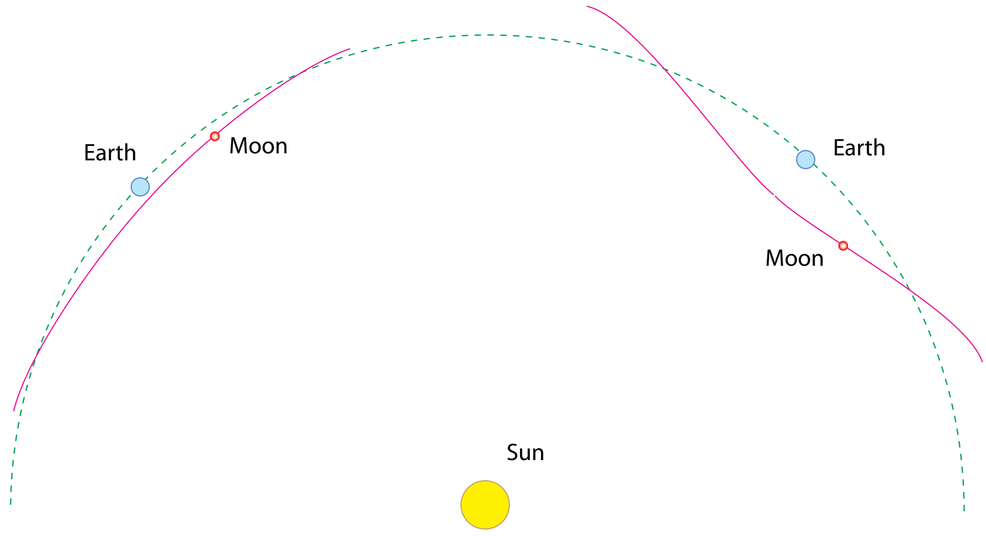


Figure 77 Which of the two Moon paths is correct?

Moon stopped rotating millions of years ago, and the Earth is on its way to doing so as well. When the Earth will have stopped rotating, the Moon will stop moving away from Earth. How far will the Moon be at that time? Afterwards however, even further in the future, the Moon will move back towards the Earth, due to the friction between the Earth–Moon system and the Sun. Even though this effect would only take place if the Sun burned forever, which is known to be false, can you explain it?

Challenge 257 ny

Challenge 258 n

▪ When you run towards the east, you *lose weight*. There are two different reasons for this: the 'centrifugal' acceleration increases and thus the force with which we are pulled down diminishes, and the Coriolis force appears, with a similar result. Can you estimate the size of the two effects?

Challenge 259 ny

▪ What is the time ratio between a stone falling through a distance l and a pendulum swinging though half a circle of radius l? (This problem is due to Galileo.) How many digits of the number π can one expect to determine in this way?

Challenge 260 n

▪ Why can a spacecraft accelerate through the *slingshot effect* when going round a planet, despite momentum conservation?

Challenge 261 n

Ref. 95

▪ The orbit of a planet around the Sun has many interesting properties. What is the hodograph of the orbit? What is the hodograph for parabolic and hyperbolic orbits?

Challenge 262 n

▪ A simple, but difficult question: if all bodies attract each other, why don't or didn't all stars fall towards each other?

Challenge 263 n

Ref. 110

▪ The acceleration g due to gravity at a depth of 3000 km is 10.05 m/s^2, over 2 % higher than at the surface of the Earth. How is this possible? Also on the Tibetan plateau, g is higher than the sea level value of 9.81 m/s^2, even though the plateau is more distant from the centre of the Earth than sea level is. How is this possible?

Challenge 264 n

▪ When the Moon circles the Sun, does its path have sections *concave* towards the Sun, as shown in the right part of Figure 77, or not, as shown on the left part? (Independently of this issue, both paths in the drawing hide that the Moon path does *not* lie in the same plane as the path of the Earth around the Sun.)

Challenge 265 n

Figure 78 The analemma over Delphi, between January and December 2002 (© Anthony Ayiomamitis)

▪ You can prove that objects *attract each other* (and that they are not attracted by the Earth only) with a simple experiment which everybody can perform at home, as described on the http://www.fourmilab.ch/gravitation/foobar/ website.

▪ It is instructive to calculate the *escape velocity* of the Earth, i.e. that velocity with which a body must be thrown so that it never falls back. It turns out to be 11 km/s. What is the escape velocity for the solar system? By the way, the escape velocity of our galaxy is 129 km/s. What would happen if a planet or a system were so heavy that its escape velocity would be larger than the speed of light? Challenge 266 n

▪ Can gravity produce repulsion? What happens to a small test body on the inside of a large C-shaped mass? Is it pushed towards the centre of mass? Challenge 267 ny

▪ For bodies of irregular shape, the centre of gravity of a body is *not* the same as the centre of mass. Are you able to confirm this? (Hint: find and use the simplest example possible.) Challenge 268 n

▪ The *shape* of the Earth is not a sphere. As a consequence, a plumb line usually does not point to the centre of the Earth. What is the largest deviation in degrees? Ref. 111 Challenge 269 ny

▪ What is the largest asteroid one can escape from by jumping? Challenge 270 n

▪ If you look at the sky every day at 6 a.m., the Sun's position varies during the year. The result of photographing the Sun on the same film is shown in Figure 78. The curve, called the analemma, is due to the inclination of the Earth's axis, and due to the elliptical shape of the path around the Sun. The shape of the analemma is also built into high quality sundials. The top and the (hidden) bottom points correspond to the solstices.

▪ The constellation in which the Sun stands at noon (at the centre of the time zone) is supposedly called the 'zodiacal sign' of that day. Astrologers say there are twelve of them, namely Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpius, Sagittarius, Capricornus, Aquarius and Pisces and that each takes (quite precisely) a twelfth of a year or a twelfth of the ecliptic. Any check with a calendar shows that at present, the midday Sun is never in the zodiacal sign during the days usually connected to it. The relation has Page 97 shifted by about a month since it was defined, due to the precession of the Earth's axis. A check with a map of the star sky shows that the twelve constellations do not have the same length and that on the ecliptic there are fourteen of them, not twelve. There is *Ophiuchus*, the snake constellation, between Scorpius and Sagittarius, and *Cetus*, the whale, between Aquarius and Pisces. In fact, not a single astronomical statement about zodiacal signs is Ref. 112 correct. To put it clearly, astrology, in contrast to its name, is *not* about stars. (In some languages, the term for 'crook' is derived from the word 'astrologer'.)

▪ The gravitational acceleration for a particle inside a spherical shell is zero. The vanishing of gravity in this case is independent of the particle shape and its position, and independent of the thickness of the shell.* Can you find the Challenge 271 ny argument using Figure 79? This works only because of the $1/r^2$ dependence of gravity. Can you show that the result does not hold for non-spherical shells? Note that the vanishing of gravity inside a spherical shell usually does not hold if other matter is found outside the shell. How could one Challenge 272 ny eliminate the effects of outside matter?

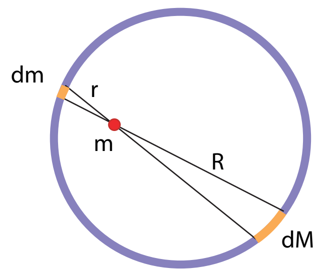


Figure 79 The vanishing of gravitational force inside a spherical shell of matter

▪ There is no planet X, i.e. no tenth planet in our solar Ref. 113 system outside Neptune and Pluto. But there are many small objects beyond them, in the so-called Kuiper belt and Oort cloud. Sometimes they change trajectory due to the attraction of a planet: that is the birth of a new comet.

▪ In astronomy new examples of motion are regularly discovered even in the present century. Sometimes there are also false alarms. One example was the alleged fall of *mini comets* on the Earth. They were supposedly made of a few dozens of kilograms of ice and Ref. 114 hitting the Earth every few seconds. It is now known not to happen. On the other hand, it is known that many tons of asteroids fall on the Earth every day, in the form of tiny particles. By the way, discovering objects hitting the Earth is not at all easy. Astronomers like to point out that an asteroid as large as the one which led to the extinction of the dinosaurs could hit the Earth without any astronomer noticing beforehand, if the direction is slightly unusual, such as from the south, where few telescopes are located.

▪ Universal gravity allows only elliptical, parabolic or hyperbolic orbits. It is impossible for a small object approaching a large one to be captured. At least, that is what we have learned so far. Nevertheless, all astronomy books tell stories of capture in our solar system, Challenge 273 ny e.g. about several outer satellites of Saturn. How is this possible?

▪ How would a tunnel have to be shaped in order that a stone falls through it without touching the walls? (Assume constant density.) If the Earth would not rotate, the tunnel would be a straight line through its centre, and the stone would fall down and up again,

* This is a small example from the beautiful text by Mark P. Silverman, *And Yet It Moves: Strange Systems and Subtle Questions in Physics*, Cambridge University Press, 1993. It is a treasure chest for anybody interested in the details of physics.

in a oscillating motion. For a rotating Earth, the problem is much more difficult. What is the shape when the tunnel starts at the Equator? Challenge 274 n

▪ The International Space Station circles the Earth at an altitude of about 380 km every 90 minutes. You can see where it is from the website htpp://www.heavens-above.com. By the way, whenever it is just above the horizon, the station is the third brightest object in the night sky, superseded only by the Moon and Venus. Have a look at it. Challenge 275 e

▪ Is it true that the centre of mass of the solar system is always inside the Sun? Challenge 276 ny

▪ All points on the Earth do not receive the same number of daylight hours during a year. The effects are difficult to spot, though. Can you find one? Challenge 277 d

▪ Can the phase of the Moon have a measurable effect on the human body? What about the tidal effects of the Moon? Challenge 278 ny

▪ There is an important difference between the heliocentric system and the old idea that all planets turn around the Earth. The heliocentric system states that certain planets, such as Mars or Venus, can be *between* the Earth and the Sun at certain times, and *behind* the Sun at other times. In contrast, the geocentric system states that they are always in between. Why did such an important difference not invalidate the geocentric system right away? Challenge 279 ny

▪ The strangest reformulation of the description of motion given by $m\mathbf{a} = \nabla U$ is the almost absurd looking equation Ref. 115

$$\nabla v = \mathrm{d}\mathbf{v}/\mathrm{d}s \tag{48}$$

where s is the motion path length. It is called the *ray form* of Newton's equation of motion. Can you find an example of its application? Challenge 280 ny

▪ Seen from Neptune, the size of the Sun is the same as that of Jupiter seen from the Earth at the time of its closest approach. True? Challenge 281 n

▪ What is gravity? This is not a simple question. In 1747, Georges-Louis Lesage proposed an explanation for the $1/r^2$ dependence. He argued that the world is full of small particles – he called them 'corpuscules ultra-mondains' – flying around randomly and hitting all objects. Single objects do not feel the hits, since they are hit continuously and randomly from all directions. But when two objects are near each other, they produce shadows for part of the flux to the other body, resulting in an attraction. Can you show that such an attraction has a $1/r^2$ dependence? Ref. 116 Challenge 282 ny

However, Lesage's proposal has a number of problems. The argument only works if the collisions are inelastic. (Why?) However, that would mean that all bodies would heat up with time, as Jean-Marc Lévy-Leblond explains. Ref. 2

There are even more problems with the idea of Lesage. First, a moving body in free space would be hit by more or faster particles in the front than in the back; as a result, the body should be decelerated. Second, gravity would depend on size, but in a strange way. In particular, three bodies aligned on a line should *not* produce shadows, as no such shadows are observed.

Despite all the criticisms, this famous idea has resurfaced in physics regularly ever since, even though such particles have never been found. Only in the third part of our mountain ascent will we settle the issue.

▪ For which bodies does gravity decrease when approaching them? Challenge 283 ny

▪ Could one put a satellite into orbit using a cannon? Does the answer depend on the direction in which one shoots? Challenge 284 ny

▪ Two computer users share experiences. 'I threw my Pentium III and Pentium IV out of the window.' 'And?' 'The Pentium III was faster.'

Challenge 285 n ▪ How often does the Earth rise and fall when seen from the Moon? Does the Earth show phases?

Challenge 286 ny ▪ What is the weight of the Moon? How does it compare to the weight of the Alps?

▪ Due to the slightly flattened shape of the Earth, the source of the Mississippi is about 20 km nearer to the centre of the Earth than its mouth; the water effectively runs uphill. Challenge 287 n How can this be?

▪ If a star is made of high density material, the orbital speed of a planet circling it close Challenge 288 n by could be larger than the speed of light. How does nature avoid this strange possibility?

▪ What will happen to the solar system in the future? This question is surprisingly hard to answer. The main expert of this topic, US physicist Gerald Sussman, simulated a few hundred million years of evolution on specially built computers, following only Ref. 117 the planets, without taking into account the smaller objects. He found that the planetary Page 236 orbits are stable, but that there is clear evidence of chaos in the evolution of the solar system, at a small level. The various planets influence each other in subtle and still poorly understood ways. Effects in the past are also being studied, such as the energy change of Jupiter due to its ejection of smaller asteroids from the solar system, or energy gains of Neptune. There is still a lot of research to be done in this field.

▪ One of the great open problems of the solar system is the description of planet distances discovered in 1766 by Johann Daniel Titius (1729–1796) and publicized by Johann Elert Bode (1747–1826). Titius discovered that planetary distances d from the Sun can be approximated by

$$d = a + b\,2^n \quad \text{with} \quad a = 0.4\,\mathrm{AU}\,,\; b = 0.3\,\mathrm{AU} \tag{49}$$

when distances are measured in astronomical units and n is the number of the planet. The resulting approximation is compared with observations in Table 17.

Interestingly, the last three planets, as well as the planetoids, were discovered *after* Bode's and Titius' deaths; the rule had successfully predicted Uranus' distance, as well as that of the planetoids. Despite these successes – and the failure for the last two planets – nobody has yet found a model for the formation of the planets that explains Titius' rule. The large satellites of Jupiter and of Uranus have regular spacing, but not according to the Titius–Bode rule. Explaining the rule is one of the great challenges remaining in classical mechanics. It is known that the rule must be a consequence of the formation of satellite systems. The bodies not following a fixed rule, such as the outer planets of the Sun or the outer moons of Jupiter, are believed not to be part of the original system but to have been captured later on.

▪ Around 3000 years ago, the Babylonians had measured the orbital times of the seven celestial bodies. Ordered from longest to shortest, they wrote them down in Table 18.

The Babylonians also introduced the week and the division of the day in 24 hours. The Babylonians dedicated every one of the 168 hours of the week to a celestial body, following the order of the table. They also dedicated the whole day to that celestial body that corresponds to the first hour of that day. The first day of the week was dedicated to Saturn; Challenge 289 e the present ordering of the other days of the week then follows from Table 18. This story Ref. 118 is already told by Cassius Dio (*c.* 160 to *c.* 230). Towards the end of Antiquity, the order-

Table 17 An unexplained property of nature: planet distances and the values resulting from the Titius–Bode rule

Planet	n	Predicted distance in AU	Measured distance in AU
Mercury	$-\infty$	0.4	0.4
Venus	0	0.7	0.7
Earth	1	1.0	1.0
Mars	2	1.6	1.5
Planetoids	3	2.8	2.2 to 3.2
Jupiter	4	5.2	5.2
Saturn	5	10.0	9.5
Uranus	6	19.6	19.2
Neptune	7	38.8	30.1
Pluto	8	77.2	39.5

Table 18 The orbital periods known to the Babylonians

Body	Period
Saturn	29 a
Jupiter	12 a
Mars	687 d
Sun	365 d
Venus	224 d
Mercury	88 d
Moon	29 d

ing was taken over in the Roman empire. In Germanic languages, including English, the Latin names of the celestial bodies were replaced by the corresponding Germanic gods. The order Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday and Friday is thus a consequence of both the astronomical measurements and the astrological superstitions of the ancients.

▪ In 1722, the great mathematician Leonhard Euler made a calculation mistake that led him to conclude that if a tunnel were built from one pole of the Earth to the other, a stone falling into it would arrive at the Earth's centre and then turn back up directly. Voltaire made fun of this conclusion for many years. Can you correct Euler and show that the real motion is an oscillation from one pole to the other, and can you calculate the time a pole-to-pole fall would take (assuming homogeneous density)?

Challenge 290 n

What would be the oscillation time for an arbitrary straight surface-to-surface tunnel of length l, thus *not* going from pole to pole?

Challenge 291 n

▪ Figure 80 shows a picture of a solar eclipse taken by the Russian space station Mir. It

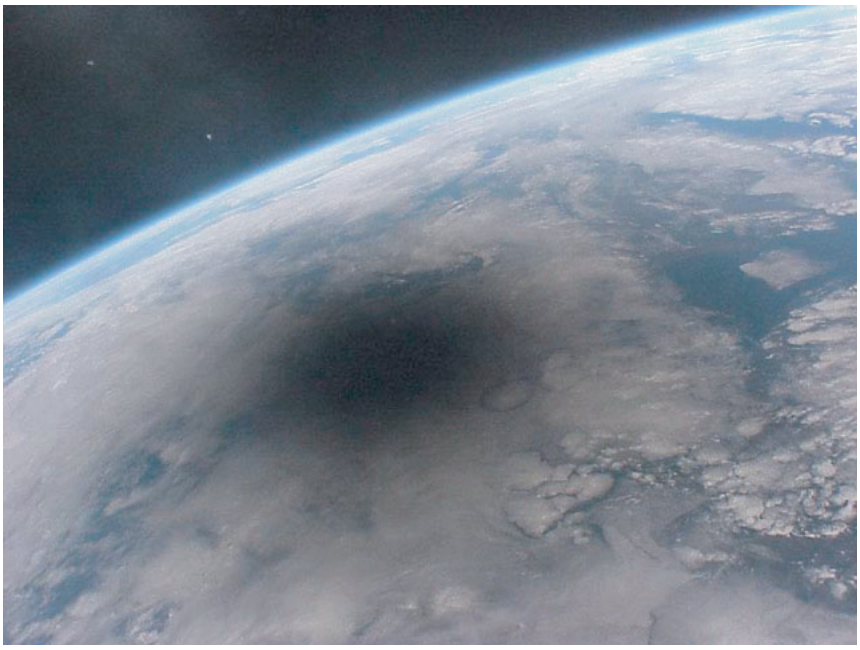


Figure 80 A solar eclipse

shows clearly that a global view of a phenomenon can be quite different from a local one.
Challenge 292 ny What is the speed of the shadow?

What is classical mechanics?

All types of motion that can be described when the mass of a body is its only permanent property form what is called *mechanics*. The same name is also given the experts studying the field. We can think of mechanics as the athletic part of physics;* both in athletics and in mechanics only lengths, times and masses are measured.

More specifically, our topic of investigation so far is called *classical* mechanics, to distinguish it from *quantum* mechanics. The main difference is that in classical physics arbitrary small values are assumed to exist, whereas this is not the case in quantum physics. The use of real numbers for observable quantities is thus central to classical physics.

Classical mechanics is often also called *Galilean physics* or *Newtonian physics*. The basis of classical mechanics, the description of motion using only space and time, is called *kinematics*. An example is the description of free fall by $z(t) = z_0 + v_0(t - t_0) - \frac{1}{2}g(t - t_0)^2$. The other, main part of classical mechanics is the description of motion as a consequence of interactions between bodies; it is called *dynamics*. An example of dynamics is the formula of universal gravity.

* This is in contrast to the actual origin of the term 'mechanics', which means 'machine science'. It derives from the Greek μηκανή, which means 'machine' and even lies at the origin of the English word 'machine' itself. Sometimes the term 'mechanics' is used for the study of motion of *solid* bodies only, excluding e.g. hydrodynamics. This use has fallen out of favour in physics in the past century.

The distinction between kinematics and dynamics can also be made in relativity, thermodynamics and electrodynamics. Even though we have not explored these fields of enquiry yet, we know that there is more to the world than gravity. A simple observation makes the point: friction. Friction cannot be due to gravity, because friction is not observed in the skies, where motion follows gravity rules only.* Moreover, on Earth, friction
Challenge 293 e is independent of gravity, as you might want to check. There must be another interaction responsible for friction. We shall study it shortly. But one issue merits a discussion right away.

Should one use force?

> The direct use of force is such a poor solution to any problem, it is generally employed only by small children and large nations.
>
> David Friedman

Everybody has to take a stand on this question, even students of physics. Indeed, many types of forces are used and observed in daily life. One speaks of muscular, gravitational, psychic, sexual, satanic, supernatural, social, political, economic and many other types of forces. Physicists see things in a simpler way. They call the different types of forces observed between objects *interactions*. The study of the details of all these interactions will show that in everyday life, they are of electrical origin.

For physicists, all change is due to motion. The term force then also gets a more restrictive definition. *(Physical) force* is defined as the *change of momentum*, i.e. as

$$\mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} . \tag{50}$$

Force is the change or flow of motion. If a force acts on a body, momentum flows into it. Indeed, momentum can be imagined as some invisible and intangible liquid. Force measures how much of this liquid flows from one body to another per unit time.

Using the Galilean definition of linear momentum $\mathbf{p} = m\mathbf{v}$, we can rewrite the definition of force as

$$\mathbf{F} = m\mathbf{a} , \tag{51}$$

where $\mathbf{F} = \mathbf{F}(t, \mathbf{x})$ is the force acting on an object of mass m and where $\mathbf{a} = \mathbf{a}(t, \mathbf{x}) = \mathrm{d}\mathbf{v}/\mathrm{d}t = \mathrm{d}^2\mathbf{x}/\mathrm{d}t^2$ is the acceleration of the same object, that is to say its change of velocity.* The expression states in precise terms that force is what changes the *velocity* of masses. The quantity is called 'force' because it corresponds in many, but *not* all aspects to muscular force. For example, the more force is used, the further a stone can be thrown.

* This is not completely correct: in the 1980s, the first case of gravitational friction was discovered: the
Page 368 emission of gravity waves. We discuss it in detail later on.

* This equation was first written down by the Swiss mathematician and physicist Leonhard Euler (1707–1783) in 1747, over 70 years after Newton's first law and 20 years after Newton's death, to whom it is usually and falsely ascribed; it was Euler, not Newton, who first understood that this definition of force is useful in *every* case of motion, whatever the appearance, be it for point particles or extended objects, and be it
Ref. 21 rigid, deformable or fluid bodies. Surprisingly and in contrast to frequently made statements, equation (51) is even correct in relativity, as shown on page 295.

However, whenever the concept of force is used, it should be remembered that *physical force is different from everyday force or everyday effort*. Effort is probably best approximated by the concept of *(physical) power*, usually abbreviated P, and defined as

$$P = \frac{\mathrm{d}W}{\mathrm{d}t} = \mathbf{F} \cdot \mathbf{v} \tag{52}$$

in which *(physical) work* W is defined as $W = \mathbf{F} \cdot \mathbf{s}$. Physical work is a form of energy, as you might want to check. Note that a man walking carrying a heavy rucksack is not doing (almost) any work at all; why then does he get tired? Work, as a form of energy, has to be taken into account when the conservation of energy is checked.

Challenge 294 ny

With the definition of work just given you can solve the following puzzles: What happens to the electricity consumption of an escalator if you walk on it instead of standing still? What is the effect of the definition of power for the salary of scientists?

Challenge 295 n

Challenge 296 n

When students in exams say that the force acting on a thrown stone is smallest at the highest point of the trajectory, it is customary to say that they are using an incorrect view, namely the so-called *Aristotelian view*, in which force is proportional to velocity. Sometimes it is even stated that they use a different concept of *state* of motion. It is then added with a tone of superiority how wrong all this is. This is a typical example of intellectual disinformation. Every student knows from riding a bicycle, from throwing a stone or from pulling objects that increased effort results in increased speed. The student is right; those theoreticians who deduce that the student has a mistaken concept of *force* are wrong. In fact, the student is just using, instead of the *physical* concept of force, the *everyday* version, namely effort. Indeed, the effort exerted by gravity on a flying stone is smallest at the highest point of the trajectory. Understanding the difference between physical force and everyday effort is the main hurdle in learning mechanics.*

Ref. 119

Often the flow of momentum, equation (50), is not recognized as the definition of force. This is mainly due to an everyday observation: there seem to be forces without any associated acceleration or momentum change, such as in a string under tension or in water of high pressure. Pushing against a tree, there is no motion, yet a force is applied. If force is momentum flow, where does the momentum go? It flows into the slight deformations of the arm and the tree. In fact, when one starts pushing and thus deforming, the associated momentum change of the molecules, the atoms, or the electrons of the two bodies can be observed. After the deformation is established, and looking at even higher magnification, one indeed finds that a continuous and equal flow of momentum is going on in both directions. By the way, the nature of this flow will be clarified in the part on quantum theory.

Since force is net momentum flow, force is needed as a separate concept only in everyday life, where it is useful in situations where net momentum flows are smaller than the total flows. At the microscopic level, momentum alone suffices for the description of motion. For example, the concept of *weight* describes the flow of momentum due to gravity. Thus we will hardly ever use the term 'weight' in the microscopic part of our adventure.

* This stepping stone is so high that many professional physicists do not really take it themselves; this is witnessed by the innumerable comments in papers which state that physical force is defined using mass, and at the same time that mass is defined using force (the latter part of the sentence being a fundamental mistake).

Through its definition the concept of force is distinguished clearly from 'mass', 'momentum', 'energy' and 'power'. But where do forces originate? In other words, which effects in nature have the capacity to accelerate bodies by pumping momentum into objects? Table 19 gives an overview.

Every example of motion, from the one that lets us choose the direction of our gaze to the one that carries a butterfly through the landscape, can be put into one of the two leftmost columns of Table 19. Physically, the two columns are separated by the following criterion: in the first class, the acceleration of a body can be in a different direction from its velocity. The second class of examples only produces accelerations exactly *opposed* to the velocity of the moving body, as seen from the frame of reference of the braking medium. Such a resisting force is called *friction*, *drag* or a *damping*. All examples in the second class are types of friction. Just check.

Challenge 298 e

Friction can be so strong that all motion of a body against its environment is made impossible. This type of friction, called *static friction* or *sticking friction*, is common and important: without it, turning the wheels of bicycles, trains or cars would have no effect. Not a single screw would stay tightened. We could neither run nor walk in a forest, as the soil would be more slippery than polished ice. In fact not only our own motion, but all *voluntary motion* of living beings is *based* on friction. The same is the case for self-moving machines. Without static friction, the propellers in ships, aeroplanes and helicopters would not be effective and the wings of aeroplanes would produce no lift to keep them in the air. In short, static friction is required whenever we want to move relative to our environment.

Once an object moves through its environment, it is hindered by another type of friction; it is called *dynamic friction* and acts between bodies in relative motion. Without it, falling bodies would always rebound to the same height without ever stopping on the floor; neither parachutes nor brakes would work; worse, we would have no memory, as we will see later on.*

As the motion examples in the second column of Table 19 include friction, in those examples macroscopic energy is not conserved; the systems are *dissipative*. In the first column, macroscopic energy is constant; the systems are *conservative*.

The first two columns can also be distinguished using a more abstract, mathematical criterion: on the left are accelerations that can be derived from a potential, on the right, decelerations that cannot. As in the case of gravitation, the description of any kind of motion is much simplified by the use of a potential: at every position in space, one needs only the single value of the potential to calculate the trajectory of an object, instead of the three values of the acceleration or the force. Moreover, the magnitude of the velocity of an object at any point can be calculated directly from energy conservation.

The processes from the second column *cannot* be described by a potential. These are the cases where we necessarily have to use force if we want to describe the motion of the

* For a general overview of the topic, from physics to economics, architecture and organizational theory, see N. ÅKERMAN, editor, *The Necessity of Friction – Nineteen Essays on a Vital Force*, Springer Verlag, 1993.

Recent research suggest that maybe in certain crystalline systems, such as tungsten bodies on silicon, under ideal conditions gliding friction can be extremely small and possibly even vanish in certain directions of motion. This so-called *superlubrication* is presently a topic of research.

Ref. 120

Table 19 Selected processes and devices changing the motion of bodies

Situations that can lead to acceleration	Situations that only lead to deceleration	Motors and actuators
piezoelectricity		
quartz under applied voltage	thermoluminescence	walking piezo tripod
gravitation		
falling	emission of gravity waves	pulley
collisions		
satellite in planet encounter	car crash	rocket motor
growth of mountains	meteorite crash	swimming of larvae
magnetic effects		
compass needle near magnet	electromagnetic braking	electromagnetic gun
magnetostriction	transformer losses	linear motor
current in wire near magnet	electric heating	galvanometer
electric effects		
rubbed comb near hair	friction between solids	electrostatic motor
bombs	fire	muscles, sperm flagella
television tube	electron microscope	Brownian motor
light		
levitating objects by light	light bath stopping atoms	(true) light mill
solar sail for satellites	light pressure inside stars	solar cell
elasticity		
bow and arrow	trouser suspenders	ultrasound motor
bent trees standing up again	pillow, air bag	bimorphs
osmosis		
water rising in trees	salt conservation of food	osmotic pendulum
electro-osmosis		tunable X-ray screening
heat & pressure		
freezing champagne bottle	surfboard water resistance	hydraulic engines
tea kettle	quicksand	steam engine
barometer	parachute	air gun, sail
earthquakes	sliding resistance	seismometer
attraction of passing trains	shock absorbers	water turbine
nuclei		
radioactivity	plunging into the Sun	supernova explosion
biology		
bamboo growth	find example! Challenge 297 ny	molecular motors

system. For example, the force F due to the wind resistance of a body is *roughly* given by

$$F = 1/2 c_{w} \rho A v^{2} \tag{53}$$

where A is the area of its cross section and v its velocity relative to the air, ρ is the density of air; the *drag coefficient* c_{w} is a pure number that depends on the shape of the moving object. (A few examples are given in Figure 81.) You may check that aerodynamic resistance Challenge 299 ny cannot be derived from a potential.*

The drag coefficient c_{w} is found experimentally to be always larger than 0.0168, which corresponds to the optimally streamlined tear shape. An aerodynamic car has a value of 0.25 to 0.3; but many sports cars share with vans values of 0.44 and higher.**

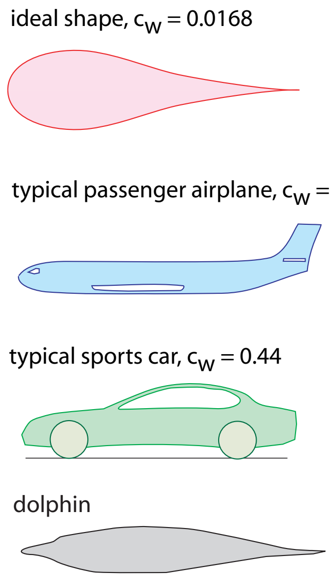


Figure 81 Shapes and air/water resistance

Wind resistance is also of importance to humans, Ref. 122 in particular in athletics. It is estimated that 100 m sprinters spend between 3 % and 6 % of their power overcoming drag. This leads to varying sprint times Challenge 301 ny t_{w} when wind of speed w is involved, related by the expression

$$\frac{t_0}{t_w} = 1.03 - 0.03 \left(1 - \frac{w t_w}{100}\right)^2 , \tag{54}$$

where the more conservative estimate of 3 % is used. An opposing wind speed of -2 m/s gives a time increase of 0.13 s, enough to change an a potential world record into an 'only' excellent result. (Are you able to deduce the c_{w} value for running humans from the for-Challenge 302 ny mula?)

Likewise, parachuting exists due to wind resistance. Can you determine how the speed Challenge 303 n of a falling body changes with time, assuming *constant* shape and drag coefficient?

In contrast, static friction has different properties. It is proportional to the force pressing Ref. 123 the two bodies together. Why? Studying the situation in more detail, sticking friction is found to be proportional to the actual contact area. It turns out that putting two solids into contact is rather like turning Switzerland upside down and putting it onto Austria; the area of contact is much smaller than the one estimated macroscopically. The important point is that actual contact area is proportional to the *normal* force. The study of what

* Such a statement about friction is correct only in three dimensions, as is the case in nature; in the case of Challenge 300 n a single dimension, a potential can *always* be found.

** Calculating drag coefficients in computers, given the shape of the body and the properties of the fluid, is one of the most difficult tasks of science; the problem is still not fully solved.

The topic of aerodynamic shapes is even more interesting for fluid bodies. They are kept together by *surface tension*. For example, surface tension keeps the hair of a wet brush together. Surface tension also determines the shape of rain drops. Experiments show that it is spherical for drops smaller than 2 mm, and that larger rain drops are *lens* shaped, with the flat part towards the bottom. The usual tear shape is *not* Ref. 121 encountered in nature; something vaguely similar to it appears during drop detachment, but *never* during drop fall.

happens in the small percentage of contact area is still a topic of research; researchers are investigating the issues using instruments such as atomic force microscopes, lateral force microscopes and triboscopes. One result of these efforts are computer hard disks with longer lifetimes, as the friction between disk and reading head is a central quantity determining the lifetime.

All examples of friction are accompanied by an increase in the temperature of the moving body. After the discovery of atoms, the reason became clear. Friction is not observed in few – e.g. 2, 3, or 4 – particle systems. Friction only appears in systems with many particles, usually millions or more. Such systems are called *dissipative*. Both the temperature changes and friction itself are due to motion of large numbers of microscopic particles against each other. This motion is not included in the Galilean description. When one does include it, friction and energy loss disappear, and potentials can then be used throughout. Positive accelerations – of microscopic magnitude – then also appear, and motion is found to be conserved. As a result, all motion is conservative on a microscopic scale. Therefore, on a microscopic scale it is possible to describe *all* motion without the concept of force.* The moral of the story is that one should use force only in one situation: in the case of friction, and only when one does not want to go into the microscopic details.**

> Et qu'avons-nous besoin de ce moteur, quand l'étude réfléchie de la nature nous prouve que le mouvement perpétuel est la première de ses lois ?***
>
> Donatien de Sade *Justine, ou les malheurs de la vertu.*

Complete states: initial conditions

> Quid sit futurum cras, fuge quaerere ...****
>
> Horace, *Odi*, lib. I, ode 9, v. 13.

We often describe the motion of a body by specifying the time dependence of its position, for example as

$$\mathbf{x}(\mathbf{t}) = \mathbf{x_0} + \mathbf{v_0}(t - t_0) + \tfrac{1}{2}\mathbf{a_0}(t - t_0)^2 + \tfrac{1}{6}\mathbf{j_0}(t - t_0)^3 + \dots \,. \quad (55)$$

* The first scientist who eliminated force from the description of nature was Heinrich Rudolf Hertz (b. 1857 Hamburg, d. 1894 Bonn), the famous discoverer of electromagnetic waves, in his textbook on mechanics, *Die Prinzipien der Mechanik*, Barth, 1894, republished by Wissenschaftliche Buchgesellschaft, Darmstadt, 1963. His idea was strongly criticized at that time; only a generation later, when quantum mechanics quietly got rid of the concept for good, did the idea become commonly accepted. (Many have speculated about the role Hertz would have played in the development of quantum mechanics and general relativity, had he not died so young.) In his book, Hertz also formulated the principle of the straightest path: particles follow geodesics. This same description is one of the pillars of general relativity, as we will see later on.

** In the case of human relations the evaluation should be somewhat more discerning. A powerful book on human violence is James Gilligan, *Violence – Our Deadly Epidemic and Its Causes*, Grosset/Putnam, 1992.

*** 'And whatfor do we need this motor, when the reasoned study of nature proves us that perpetual motion is the first of its laws?'

Ref. 54 **** 'What future be tomorrow, never ask ...' Horace is Quintus Horatius Flaccus (65–8 BCE), the great Roman poet.

The quantities with an index 0, such as the starting position $\mathbf{x_0}$, the starting velocity $\mathbf{v_0}$, etc., are called *initial conditions*. Initial conditions are necessary for any description of motion. Different physical systems have different initial conditions. Initial conditions thus specify the *individuality* of a given system. Initial conditions also allow us to distinguish the present situation of a system from that at any previous time: initial conditions specify the *changing aspects* of a system. In other words, they summarize the past of a system.

Page 38 Initial conditions are thus precisely the properties we sought for a description of the *state* of a system. To find a complete description of states we thus only need a complete description of initial conditions. It turns out that for gravitation, like for all other microscopic interactions, there is *no* need for initial acceleration $\mathbf{a_0}$, initial jerk $\mathbf{j_0}$, or higher-order initial quantities. In nature, acceleration and jerk only depend on the properties of objects and their environment; they do not depend on the past. For example, the expression $a = GM/r^2$, giving the acceleration of a small body near a large one, does not depend on the past at all, but only on the environment. The same happens for the other fundamental interactions, as we will find out shortly.

Page 67 The *complete state* of a moving mass point is thus described by specifying its position and its momentum for all instants of time. Thus we have achieved a complete description of the intrinsic properties of point objects, namely by their mass, and of their states of motion, namely by their momentum, energy, position and time. For *extended* rigid objects we also need orientation, angular velocity and angular momentum. Can you specify the necessary quantities in the case of extended elastic bodies or fluids? Challenge 304 ny

The set of all possible states of a system is given a special name: it is called the *phase space*. We will use the concept repeatedly. Like any space, it has a number of dimensions. Can you specify it for a system made of N point particles? Challenge 305 ny

However, there are situations in nature where the motion of an object depends on other characteristics than its mass; motion can depend on its colour (can you find an example?), on its temperature, and on a few other properties which we will soon discover. Challenge 306 n Can you give an example of an intrinsic property we have missed so far? Challenge 307 n And for each intrinsic property there are state variables to discover. These new properties are the basis of field of physical enquiry beyond mechanics. We must therefore conclude that we do not have a complete description of motion yet.

It is interesting to recall an older challenge and ask again: does the universe have initial conditions? Challenge 308 n Does it have a phase space? As a hint, recall that when a stone is thrown, the initial conditions summarize the effects of the thrower, his history, the way he got there etc.; in other words, initial conditions summarize the effects the environment had during the history of a system.

> An optimist is somebody who thinks that the future is uncertain.

Do surprises exist? Is the future determined?

Die Ereignisse der Zukunft *können* wir nicht aus den gegenwärtigen erschließen. Der Glaube an den Kausalnexus ist ein Aberglaube.*
Ludwig Wittgenstein, *Tractatus*, 5.1361

Freedom is the recognition of necessity.
Friedrich Engels (1820–1895)

If, after climbing a tree, we jump down, we cannot stop the jump in the middle of the trajectory; once the jump is begun, it is unavoidable and determined, like all passive motion. However, when we start moving an arm, we can stop or change its motion from a hit to a caress. Voluntary motion does not seem unavoidable or predetermined. Which of these two cases is the general one?

Challenge 309 e

Let us start with the example we can describe most precisely so far: the fall of a body. Once the potential φ acting on a particle is given and taken into account, using

$$\mathbf{a}(x) = -\nabla\varphi = -GM\mathbf{r}/r^3 , \tag{56}$$

and the state at a given time is given by initial conditions such as

$$\mathbf{x}(t_0) = x_0 \quad \text{and} \quad \mathbf{v}(t_0) = v_0 , \tag{57}$$

we then can determine the motion in advance. The complete trajectory $\mathbf{x}(t)$ can be calculated with these two pieces of information. Due to this possibility, an equation such as (56) is called an *evolution equation* for the motion of the object. (Note that the term 'evolution' has different meanings in physics and in biology.) An evolution equation always expresses the observation that not all types of change are observed in nature, but only certain specific cases. Not all imaginable sequences of events are observed, but only a limited number of them. In particular, equation (56) expresses that from one instant to the next, objects change their motion based on the potential acting on them. Thus, given an evolution equation and initial state, the whole motion of a system is *uniquely fixed*; this property of motion is often called *determinism*. Since this term is often used with different meanings, let us distinguish it carefully from several similar concepts, to avoid misunderstandings.

Motion can be deterministic and at the same time still be *unpredictable*. The latter property can have four origins: an impracticably large number of particles involved, the complexity of the evolution equations, insufficient information on initial conditions, or strange shapes of space-time. The weather is an example where the first three conditions are fulfilled at the same time.* Nevertheless, its motion is still deterministic. Near black holes all four cases apply together. We will discuss black holes in the section on general relativity. Nevertheless, near black holes, motion is still deterministic.

* We *cannot* infer the events of the future from those of the present. Superstition is nothing but belief in the causal nexus.

* For a beautiful view of clouds, see the http://www.goes.noass.gov website.

Motion can be both deterministic and time *random*, i.e. with different outcomes in similar experiments. A roulette ball's motion is deterministic, but it is also random.* As we will see later, quantum-mechanical situations fall into this category, as do all examples of irreversible motion, such as an drop of ink spreading in clear water. In all such cases the randomness and the irreproducibility are only apparent; they disappear when the description of states and initial conditions in the microscopic domain are included. In short, determinism does not contradict *(macroscopic) irreversibility*. However, on the microscopic scale, deterministic motion is always reversible.

A final concept to be distinguished from determinism is *acausality*. Causality is the requirement that cause must precede the effect. This is trivial in Galilean physics, but becomes of importance in special relativity, where causality implies that the speed of light is a limit for the spreading of effects. Indeed, it seems impossible to have deterministic motion (of matter and energy) which is *acausal*, i.e. faster than light. Can you confirm
Challenge 310 n this? This topic will be deepened in the section on special relativity.

Saying that motion is 'deterministic' means that it is fixed in the future *and also in the past*. It is sometimes stated that predictions of *future* observations are the crucial test for a successful description of nature. Due to our often impressive ability to influence the future, this is not necessarily a good test. Any theory must, first of all, describe *past* observations correctly. It is our lack of freedom to change the past that results in our lack of choice in the description of nature that is so central to physics. In this sense, the term 'initial condition' is an unfortunate choice, because it automatically leads us to search for the initial condition of the universe and to look there for answers to questions that can be answered without that knowledge. The central ingredient of a deterministic description is that all motion can be reduced to an evolution equation plus one specific state. This state can be either initial, intermediate, or final. Deterministic motion is uniquely specified into the past and into the future.

To get a clear concept of determinism, it is useful to remind oneself why the concept of 'time' is introduced in our description of the world. We introduce time because we observe first that we are able to define sequences among observations, and second, that unrestricted change is impossible. This is in contrast to movies, where one person can walk through a door and exit into another continent or another century. In nature we do not observe metamorphoses, such as people changing into toasters or dogs into toothbrushes. We are able to introduce 'time' only because the sequential changes we observe
Challenge 311 n are extremely restricted. If nature were not reproducible, time could not be used. In short, determinism expresses the observation that *sequential changes are restricted to a single possibility*.

Since determinism is connected to the use of the concept of time, new questions arise whenever the concept of time changes, as happens in special relativity, in general relativity and in theoretical high energy physics. There is a lot of fun ahead.

In summary, every description of nature that uses the concept of time, such as that of everyday life, that of classical physics and that of quantum mechanics, is intrinsically and

* Mathematicians have developed a large number of tests to determine whether a collection of numbers may be called *random*; roulette results pass all these tests – in honest casinos only, however. Such tests typically check the equal distribution of numbers, of pairs of numbers, of triples of numbers, etc. Other tests are the
Ref. 124 χ^2 test, the Monte Carlo test(s), and the gorilla test.

inescapably deterministic, since it connects observations of the past and the future, *eliminating* alternatives. In short, *the use of time implies determinism, and vice versa.* When drawing metaphysical conclusions, as is so popular nowadays when discussing quantum Page 775 theory, one should never forget this connection. Whoever uses clocks but denies determinism is nurturing a split personality!*

The idea that motion is determined often produces fear, because we are taught to associate determinism with lack of freedom. On the other hand, we do experience freedom in our actions and call it *free will.* We know that it is necessary for our creativity and for our happiness. Therefore it seems that determinism is opposed to happiness.

But what is free will precisely? Much ink has been consumed trying to find a precise definition. One can try to define free will as the arbitrariness of the choice of initial conditions. However, initial conditions must themselves result from the evolution equations, so that there is in fact no freedom in their choice. One can try to define free will from the idea of unpredictability, or from similar properties, such as uncomputability. But these definitions face the same simple problem: whatever the definition, there is *no way* to prove experimentally that an action was performed freely. The possible definitions are useless. In short, free will *cannot* be observed. (Psychologists also have a lot of their own data to underline this, but that is another topic.)

No process that is *gradual* – in contrast to *sudden* – can be due to free will; gradual processes are described by time and are deterministic. In this sense, the question about free will becomes one about the existence of sudden changes in nature. This will be a recurring topic in the rest of this walk. Does nature have the ability to surprise? In everyday life, nature does not. Sudden changes are not observed. Of course, we still have to investigate this question in other domains, in the very small and in the very large. Indeed, we will change our opinion several times. On the other hand, we know the result of everyday life: the concept of curiosity is based on the idea that everything discovered is useful afterwards. If nature continually surprised us, curiosity would make no sense.

Another observation speaks against surprises: in the beginning of our walk we defined time using the continuity of motion; later on we expressed this by saying that time is a consequence of the conservation of energy. Conservation is the opposite of surprise. By the way, a challenge remains: can you show that time would not be definable even if Challenge 312 ny surprises existed only *rarely*?

In summary, so far we have no evidence that surprises exist in nature. Time exists because nature is deterministic. Free will cannot be defined with the precision required by physics. Given that there are no sudden changes, there is only one consistent definition of free will: it is a *feeling*, in particular of independence of others, of independence from Ref. 125 fear and of accepting the consequences of one's actions. Free will is a feeling of satisfaction. This solves the apparent paradox; free will, being a feeling, exists as a human experience, even though all objects move without any possibility of choice. There is no contradiction.*

* That can be a lot of fun though.

* That free will is a feeling can also be confirmed by careful introspection. The idea of free will always appears *after* an action has been started. It is a beautiful experiment to sit down in a quiet environment, with the intention to make, within an unspecified number of minutes, a small gesture, such as closing a hand. If you Challenge 313 e carefully observe, in all detail, what happens inside yourself around the very moment of decision, you find either a mechanism that led to the decision, or a diffuse, unclear mist. You never find free will. Such an experiment is a beautiful way to experience deeply the wonders of the self. Experiences of this kind might

Ref. 126 Even if human action is determined, it still is authentic. So why is determinism so frightening? That is a question everybody has to ask himself. What difference does determinism imply for your life, for the actions, the choices, the responsibilities and the pleasures you encounter?* If you conclude that being determined is different from being free, you should change your life! Fear of determinism usually stems from refusal to take the world the way it is. Paradoxically, it is precisely he who insists on the existence of free will who is running away from responsibility. Challenge 314 e

> You do have the ability to surprise yourself.
> Richard Bandler and John Grinder

A strange summary about motion

> Darum kann es in der Logik auch *nie* Überraschungen geben.*
> Ludwig Wittgenstein, *Tractatus*, 6.1251

Classical mechanics describes nature in a rather simple way. *Objects* are permanent and massive entities localized in space-time. *States* are changing properties of objects, described by position in space and instant in time, by energy and momentum, and by their rotational equivalents. *Time* is the relation between events measured by a clock. *Clocks* are devices in undisturbed motion whose position can be observed. *Space* and position is the relation between objects measured by a metre stick. *Metre sticks* are devices whose shape is subdivided by some marks, fixed in an invariant and observable manner. *Motion* is change of position with time (times mass); it is determined, does not show surprises, is conserved (even in death), and is due to gravitation and other interactions.

Even though this description works rather well, it contains a circular definition. Can you spot it? Each of the two central concepts of motion is defined with the help of the other. Physicists worked for about 200 years on classical mechanics without noticing or wanting to notice the situation. Even thinkers with an interest in discrediting science did not point it out. Can an exact science be based on a circular definition? Obviously, physics has done quite well so far. Some even say the situation is unavoidable in principle. Despite these opinions, undoing this logical loop is one of the aims of the rest of our walk. To achieve it, we need to increase substantially the level of precision in our description of motion. Challenge 316 n Challenge 317 n

Whenever precision is increased, the imagination is restricted. We will discover that many types of motion that seem possible are not. Motion is limited. Nature limits speed, size, acceleration, mass, force, power and many other quantities. Continue reading only if you are prepared to exchange fantasy for precision. It will not be a loss, as you will gain something else: the workings of nature will fascinate you.

---

also be one of the origins of human spirituality, as they show the connection everybody has with the rest of nature.

Challenge 315 n * If nature's 'laws' are deterministic, are they in contrast with moral or ethical 'laws'? Can people still be held responsible for their actions?

* Hence there can *never* be surprises in logic.

Bibliography

Aiunt enim multum legendum esse, non multa.
Plinius, *Epistulae*.*

1 For a history of science in antiquity, see Lucio Russo, *La rivoluzione dimenticata*, Feltrinelli 1996. also available in several other languages. Cited on page 28.

2 A beautiful book explaining physics and its many applications in nature and technology vividly and thoroughly is Paul G. Hewitt, John Suchocki & Leslie A. Hewitt, *Conceptual Physical Science*, Bejamin/Cummings, 1999.

A book famous for its passion for curiosity is Richard P. Feynman, R.B. Leighton & M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, 1977.

A lot on motion can be learned from quiz books. One of the best is the well-structured collection of beautiful problems that require no mathematics written by Jean-Marc Lévy-Leblond, *La physique en questions – mécanique*, Vuibert, 1998.

Another excellent quiz collection is Yakov Perelman, *Oh, la physique*, Dunod, 2000, a translation from the Russian original.

A good problem book is W.G. Rees, *Physics by Example: 200 Problems and Solutions*, Cambridge University Press, 1994.

A good history of physical ideas is given in the excellent text by David Park, *The How and the Why*, Princeton University Press, 1988. Cited on pages 86, 130, and 199.

3 A well-known principle in the social sciences states that given a question, for every possible answer, however weird it may seem, there is somebody – and often a whole group – who holds it as his opinion. One just has to go through literature (or the internet) to find out.

About group behaviour in general, see R. Axelrod, *The Evolution of Cooperation*, Harper Collins, 1984. The propagation and acceptance of ideas, such as those of physics, are also an example of human cooperation, with all its potential dangers and weaknesses. Cited on page 30.

4 All the known texts by Parmenides and Heraclitos can be found in Jean-Paul Dumont, Les écoles présocratiques, Folio-Gallimard, 1988. Views about the non-existence of motion have also been put forward by much more modern and much more contemptible authors, such as in 1710 by Berkeley. Cited on page 30.

Page 630

5 An example of people worried by Zeno is given by William McLaughlin, Resolving Zeno's paradoxes, *Scientific American* pp. 66–71, November 1994. The actual argument was not about a hand slapping a face, but about an arrow hitting the target. See also reference Ref. 41. Cited on page 30.

6 The most famous text is Jearl Walker, *The Flying Circus of Physics*, Wiley, 1975. For more interesting physical effects in everyday life, see Erwein Flachsel, *Hundertfünfzig Physikrätsel*, Ernst Klett Verlag, 1985. The book also covers several clock puzzles, in puzzle numbers 126 to 128. Cited on page 31.

* 'Read much, but not anything.' Ep. 7, 9, 15. Gaius Plinius Secundus (b. 23/4 Novum Comum, d. 79 Vesuvius eruption), Roman writer, especially famous for his large, mainly scientific work *Historia naturalis*, which has been translated and read for almost 2000 years.

7 A concise and informative introduction into the history of classical physics is given in the first chapter of the book by F.K. Richtmeyer, E.H. Kennard & J.N. Cooper, *Introduction to Modern Physics*, McGraw–Hill, 1969. Cited on page 32.

8 A good overview over the arguments used to prove the existence of god from motion is given by Michael Buckley, *Motion and Motion's God*, Princeton University Press, 1971. The intensity of the battles waged around these failed attempts is one of the tragicomic chapters of history. Cited on page 33.

9 Thomas Aquinas, *Summa Theologiae* or *Summa Theologica*, 1265–1273, online in Latin at http://www.newadvent.org/summa, in English on several other servers. Cited on page 33.

10 See e.g. the fascinating text by David G. Chandler, *The Campaigns Of Napoleon - The Mind And Method Of History's Greatest Soldier*, Macmillan, 1966. Cited on page 33.

11 Richard Marcus, *American Roulette*, St. Martin's Press, 2003, a thriller and a true story. Cited on page 33.

12 A good and funny book on behaviour change is the well-known text by R. Bandler, *Using Your Brain for a Change*, Real People Press, 1985. See also Richard Bandler & John Grinder, Frogs into princes – Neuro Linguistic Programming, Eden Grove Editions, 1990. Cited on pages 33 and 40.

13 A beautiful book about the mechanisms of human growth from the original cell to full size is Lewis Wolpert, *The Triumph of the Embryo*, Oxford University Press, 1991. Cited on page 33.

14 On the topic of grace and poise, see e.g. the numerous books on the Alexander technique, such as M. Gelb, *Body Learning – an Introduction to the Alexander Technique*, Aurum Press, 1981, and Richard Brennan, *Introduction to the Alexander Technique*, Little Brown and Company, 1996. Among others, the idea of the Alexander technique is to return to the situation that the muscle groups for sustention and those for motion are used only for their respective function, and not vice versa. Any unnecessary muscle tension, such as neck stiffness, is a waste of energy due to the use of sustention muscles for movement and of motion muscles for sustention. The technique teaches the way to return to the natural use of muscles.

Motion of animals was discussed extensively already in the seventeenth century by G. Borelli, *De motu animalium*, 1680. An example of a more modern approach is J.J. Collins & I. Stewart, Hexapodal gaits and coupled nonlinear oscillator models, *Biological Cybernetics* **68**, pp. 287–298, 1993. See also I. Stewart & M. Golubitsky, *Fearful Symmetry*, Blackwell, 1992. Cited on pages 35 and 88.

15 The results on the development of children mentioned here and in the following have been drawn mainly from the studies initiated by Jean Piaget; for more details on child development, see the intermezzo following this chapter, on page 587. At http://www.piaget.org you can find the website maintained by the Jean Piaget Society. Cited on pages 36, 42, and 44.

16 The reptilian brain (eat? flee? ignore?), also called the R-complex, includes the brain stem, the cerebellum, the basal ganglia and the thalamus; the old mammalian (emotions) brain, also called the limbic system, contains the amygdala, the hypothalamus and the hippocampus; the human (and primate) (rational) brain, called the neocortex, consists of the famous grey matter. More details can be found in the text by ... Cited on page 36.

17 The lower left corner movie can be reproduced on a computer after typing the following lines in the Mathematica software package: Cited on page 37.


```
    « Graphics'Animation'
Nxpixels=72; Nypixels=54; Nframes=Nxpixels 4/3;
```

```
Nxwind=Round[Nxpixels/4]; Nywind=Round[Nypixels/3];
front=Table[Round[Random[]],{y,1,Nypixels},{x,1,Nxpixels}];
back =Table[Round[Random[]],{y,1,Nypixels},{x,1,Nxpixels}];
frame=Table[front,{nf,1,Nframes}];
Do[ If[ x>n-Nxwind && x<n && y>Nywind && y<2Nywind,
  frame[[n,y,x]]=back[[y,x-n]] ],
     {x,1,Nxpixels}, {y,1,Nypixels}, {n,1,Nframes}];
film=Table[ListDensityPlot[frame[[nf]], Mesh-> False,
  Frame-> False, AspectRatio-> N[Nypixels/Nxpixels],
  DisplayFunction-> Identity],   {nf,1,Nframes}]
ShowAnimation[film]
```

But our motion detection system is much more powerful than the example shown in the lower left corners. The following, different movie makes the point.

```
    « Graphics‘Animation‘
Nxpixels=72; Nypixels=54; Nframes=Nxpixels 4/3;
Nxwind=Round[Nxpixels/4]; Nywind=Round[Nypixels/3];
front=Table[Round[Random[]],{y,1,Nypixels},{x,1,Nxpixels}];
back =Table[Round[Random[]],{y,1,Nypixels},{x,1,Nxpixels}];
frame=Table[front,{nf,1,Nframes}];
Do[ If[ x>n-Nxwind && x<n && y>Nywind && y<2Nywind,
  frame[[n,y,x]]=back[[y,x]] ],
     {x,1,Nxpixels}, {y,1,Nypixels}, {n,1,Nframes}];
film=Table[ListDensityPlot[frame[[nf]], Mesh-> False,
  Frame-> False, AspectRatio-> N[Nypixels/Nxpixels],
  DisplayFunction-> Identity],   {nf,1,Nframes}]
ShowAnimation[film]
```

Similar experiments, e.g. using randomly changing random patterns, show that the eye perceives motion even in cases where all Fourier components of the image are practically zero; such image motion is called *drift-balanced* or *non-Fourier* motion. Several examples are presented in J. Zanker, Modelling human motion perception I: classical stimuli, *Naturwissenschaften* **81**, pp. 156–163, 1994, and J. Zanker, Modelling human motion perception II: beyond Fourier motion stimuli, *Naturwissenschaften* **81**, pp. 200–209, 1994.

18 An introduction into perception research is E. Bruce Goldstein, *Perception*, Books/Cole, 5th edition, 1998. Cited on pages 33 and 37.

19 All fragments from Heraclitus are from John Mansley Robinson, *An Introduction to Early Greek Philosophy*, Houghton Muffin 1968, chapter 5. Cited on pages 37 and 174.

20 An introduction to Newton the alchemist are the two books by Betty Jo Teeter Dobbs, *The Foundations of Newton's Alchemy*, Cambridge University Press, 1983, and *The Janus Face of Genius*, Cambridge University Press, 1992. Newton is found to be a sort of highly intellectual magician, desperately looking for examples of processes where gods interact with the material world. An intense but tragic tale. A good overview is provided by R.G. Keesing, Essay Review: Newton's Alchemy, *Contemporary Physics* **36**, pp. 117–119, 1995.

Newton's infantile theology, typical for god seekers who grew up without a father, can be found in the many books summarizing the letter exchanges between Clarke, his secretary, and Leibniz, Newton's rival for fame. Cited on page 41.

21 An introduction to the story of classical mechanics, which also destroys a few of the myths surrounding it – such as the idea that Newton could solve differential equations or that he introduced the expression $F = ma$ – is given by Clifford A. Truesdell, *Essays in the History of Mechanics*, Springer, 1968. Cited on pages 41, 116, and 134.

22 Almost all textbooks, both for schools and for university start with the definition of space and time. Even otherwise excellent relativity textbooks cannot avoid this habit, even those which introduce the now standard k-calculus (which is in fact the approach mentioned here). Cited on page 41.

23 C. Liu, Z. Dutton, C.H. Behroozi & L.V. Han, Observation of coherent optical storage in an atomic medium using halted light pulses, *Nature* **409**, pp. 490–493, 2001. There is also a comment of the paper by E.A. Cornell, Stopping light in its track, **409**, pp. 461–462, 2001. However, despite the claim, the light pulses of course have *not* been halted. Can you give at least two reasons without even reading the paper, and maybe a third after reading it? Challenge 318 n

The work was an improvement of the previous experiment where a group velocity of light of 17 m/s had been achieved, in an ultracold gas of sodium atoms, at nanokelvin temperatures. This was reported by Lene Vestergaard Hau, S.E. Harris, Zachary Dutton & Cyrus H. Bertozzi, Light speed reduction to 17 meters per second in an ultracold atomic gas, *Nature* **397**, pp. 594–598, 1999. Cited on pages 43 and 279.

24 Rainer Flindt, *Biologie in Zahlen – Eine Datensammlung in Tabellen mit über 10.000 Einzelwerten*, Spektrum Akademischer Verlag, 2000. Cited on page 43.

25 Two jets with that speed have been observed by I.F. Mirabel & L.F. Rodríguez, A superluminal source in the Galaxy, *Nature* **371**, pp. 46–48, 1994, as well as the comments on p. 18. Cited on page 43.

26 An introduction to the sense of time as a result of clocks in the brain is found in R.B. Ivry & R. Spencer, The neural representation of time, *Current Opinion in Neurobiology* **14**, pp. 225–232, 2004. The chemical clocks in our body are described in John D. Palmer, *The Living Clock*, Oxford University Press, 2002, or in A. Ahlgren & F. Halberg, *Cycles of Nature: An Introduction to Biological Rhythms*, National Science Teachers Association, 1990. See also the http://www.msi.umn.edu/~halberg/introd/ website. Cited on page 46.

27 This has been shown among others by the work of Anna Wierzbicka mentioned in more detail in the Intermezzo following this chapter, on page 596. The passionate best seller by the Chomskian author Steven Pinker, *The Language Instinct – How the Mind Creates Language*, Harper Perennial, 1994, also discusses issues related to this matter, refuting amongst others on page 63 the often repeated false statement that the *Hopi* language is an exception. Cited on page 46.

28 Aristotle rejects the idea of the flow of time in chapter IV of his *Physics*. See the full text on the http://classics.mit.edu/Aristotle/physics.4.iv.html website. Cited on page 47.

29 Perhaps the most informative of the books about the 'arrow of time' is Hans Dieter Zeh, *The Physical Basis of the Direction of Time*, Springer Verlag, 4th edition, 2001. It is still the best book on the topic. Most other texts – have a look on the internet – lack clarity of ideas.

A typical conference proceeding is J.J. Halliwell, J. Pérez–Mercader & Wojciech H. Zurek, *Physical Origins of Time Asymmetry*, Cambridge University Press, 1994. Cited on page 47.

30 On the issue of absolute and relative motion there are many books about few issues. Examples are John Barbour, *Absolute or Relative Motion? Vol. 1: A Study from the Machian Point of View of the Discovery and the Structure of Spacetime Theories*, Cambridge University Press, 1989, John Barbour, *Absolute or Relative Motion? Vol. 2: The Deep Structure of General Relativity*, Oxford University Press, 2005, or John Earman, *World Enough and Spacetime: Absolute vs Relational Theories of Spacetime*, MIT Press, 1989. Cited on page 51.

31 R. Dougherty & M. Foreman, Banach–Tarski decompositions using sets with the property of Baire, *Journal of the American Mathematical Society* 7, pp. 75–124, 1994. See also Alan L.T. Paterson, *Amenability*, American Mathematical Society, 1998, and Robert M. French, The Banach–Tarski theorem, *The Mathematical Intelligencer* **10**, pp. 21–28, 1998. Finally, there are the books by Bernard R. Gelbaum & John M.H. Olmsted, *Counterexamples in Analysis*, Holden–Day, 1964, and their *Theorems and Counterexamples in Mathematics*, Springer, 1993. Cited on page 54.

32 The beautiful but not easy text is Steve Wagon, *The Banach Tarski Paradox*, Cambridge University Press, 1993. Cited on pages 54 and 1136.

33 About the shapes of salt water bacteria, see the corresponding section in the interesting book by Bernard Dixon, *Power Unseen – How Microbes Rule the World*, W.H. Freeman, 1994. The book has about 80 sections, in which as many microorganisms are vividly presented. Cited on page 54.

34 The smallest distances are probed in particle accelerators; the distance can be determined from the energy of the particle beam. In 1996, the value of 10^{-19} m was taken from the experiments described in F. Abe & al., Measurement of dijet angular distributions by the collider detector at Fermilab, *Physical Review Letters* 77, pp. 5336–5341, 1996. Cited on page 57.

35 Alexander K. Dewdney, The Planiverse – Computer Contact with a Two-dimensional World, Poseidon Books/Simon & Schuster, 1984. Several other fiction authors had explored the option of a two-dimensional universe before, always answering, incorrectly, in the affirmative. Cited on page 60.

36 There is a whole story behind the variations of g. It can be found in Chuji Tsuboi, Gravity, Allen & Unwin, 1979, or in Wolfgang Torge, Gravimetry, de Gruyter, 1989, or in Milan Burša & Karel Pěč, *The Gravity Field and the Dynamics of the Earth*, Springer, 1993. The variation of the height of the soil by around 0.3 m due to the Moon is one of the interesting effects found by these investigations. Cited on pages 61 and 121.

37 Andrea Frova, *La fisica sotto il naso – 44 pezzi facili*, Biblioteca Universale Rizzoli, Milano, 2001. Cited on page 62.

38 The study of shooting shit and its mechanisms is a part of modern biology. The reason that caterpillars do this was determined by M. Weiss, Good housekeeping: why do shelter-dwelling caterpillars fling their frass?, *Ecology Letters* **6**, pp. 361–370, 2003, who also gives the present record of 1.5 m for the 24 mg pellets of *Epargyreus clarus*. The picture of the flying frass is from S. Caveney, H. McLean & D. Surry, Faecal firing in a skipper caterpillar is pressure-driven, *The Journal of Experimental Biology* **201**, pp. 121–133, 1998. Cited on page 62.

39 This was discussed in the *Frankfurter Allgemeine Zeitung*, 2nd of August, 1997, at the time of the world athletics championship. The values are for the fastest part of a 100 m sprinter; the exact values cited were called the running speed world records in 1997, and were given as 12.048 m/s = 43.372 km/h by Ben Johnson for men, and 10.99 m/s = 39.56 km/h for women. Cited on page 63.

40 Long jump data and literature can be found in three articles all entitled Is a good long jumper a good high jumper?, in the *American Journal of Physics* **69**, pp. 104–105, 2001. In particular, world class long jumpers run at 9.35 ± 0.15 m/s, with vertical take-off speeds of 3.35 ± 0.15 m/s, giving take-off angles of about (only) 20°. A new technique for achieving higher take-off angles would allow the world long jump record to increase dramatically. Cited on page 63.

41 The arguments of Zeno can be found in Aristotle, *Physics*, VI, 9. It can be found translated in almost any language. The http://classics.mit.edu/Aristotle/physics.6.vi.html website

provides an online version in English. Cited on pages 65 and 145.

42 Professor to student: What is the derivative of velocity? Acceleration! What is the derivative of acceleration? I don't know. *Jerk*! The fourth, fifth and sixth derivatives of position are sometimes called *snap*, *crackle* and *pop*. Cited on page 66.

43 Etymology can be a fascinating topic, e.g. when it discovers the origin of the German word 'Weib' ('woman', related to English 'wife'). It was discovered, via a few Tocharian texts – an extinct Indo-European language from a region inside modern China – to mean originally 'shame'. It was used for the female genital region in an expression meaning 'place of shame'. With time, this expression became to mean 'woman' in general, while being shortened to the second term only. This story was discovered by the German linguist Klaus T. Schmidt; it explains in particular why the word is not feminine but neutral, i.e. why it uses the article 'das' instead of 'die'. Julia Simon, private communication.

Etymology can also be simple and plain fun, for example when one discovers that 'testimony' and 'testicle' have the same origin; indeed in Latin the same word 'testis' was used for both concepts. Cited on pages 66 and 73.

44 An overview of the latest developments is given by J.T. Armstrong, D.J. Hunter, K.J. Johnston & D. Mozurkewich, Stellar optical interferometry in the 1990s, *Physics Today* pp. 42–49, May 1995. More than 100 stellar diameters have been measured in 1995. Several dedicated powerful instruments are being planned. Cited on page 68.

45 A good biology textbook on growth is ... Cited on page 69.

46 This is discussed for example in C.L. Stong, The amateur scientist – how to supply electric power to something which is turning, *Scientific American* pp. 120–125, December 1975. It also discusses how to make a still picture of something rotating simply using a few prisms, the so-called *Dove prisms*. Other examples of attaching something to a rotating body are given by E. Rieflin, Some mechanisms related to Dirac's strings, *American Journal of Physics* 47, pp. 379–381, 1979. Cited on page 69.

47 James A. Young, Tumbleweed, *Scientific American* **264**, pp. 82–87, March 1991. The tumbleweed is in fact quite rare, except in in Hollywood westerns, where all directors feel obliged to give it a special appearance. Cited on page 69.

48 The first experiments to prove the rotation of the flagella were by M. Silverman & M.I. Simon, Flagellar rotation and the mechanism of bacterial motility, *Nature* **249**, pp. 73–74, 1974. For some pretty pictures of the molecules involved, see K. Namba, A biological molecular machine: bacterial flagellar motor and filament, *Wear* **168**, pp. 189–193, 1993. The present recordspeed of rotation, 1700 rotations per second, is reported by Y. Magariyama, S. Sugiyama, K. Muramoto, Y. Maekawa, I. Kawagishi, Y. Imae & S. Kudo, Very fast flagellar rotation, *Nature* **371**, p. 752, 1994.

More on bacteria can be learned from David Dusenbery, *Life at a Small Scale*, Scientific American Library, 1996. Cited on page 70.

49 On shadows, see the agreeable popular text by Roberto Casati, *Alla scoperta dell'ombra – Da Platone a Galileo la storia di un enigma che ha affascinato le grandi menti dell'umanità*, Oscar Mondadori, 2000, and his websites located at http://www.shadowmill.com and http://roberto.casati.free.fr/casati/roberto.htm. Cited on page 70.

50 There is also the beautiful book by Penelope Farrant, *Colour in Nature*, Blandford, 1997. Cited on page 71.

51 The laws of cartoon physics can be found in various places on the internet. Cited on page 71.

52 For the curious, an overview of the illusions used in the cinema and in television, which

lead to some of the strange behaviour of images mentioned above, is given in Bernard Wilkie, *The Technique of Special Effects in Television*, Focal Press, 1993, and his other books, or in the *Cinefex* magazine. Cited on page 71.

53 Aetius, Opinions, I, XXIII, 3. See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 426, 1991. Cited on page 72.

54 Giuseppe Fumagalli, *Chi l'ha detto?*, Hoepli, 1983. The sentence is also the motto of the cities of the Hansa. Cited on pages 72, 139, and 158.

55 For the role and chemistry of adenosine triphosphate (ATP) in cells and in living beings, see any chemistry book, or search the internet. The uncovering of the mechanisms around ATP has led to chemistry Nobel Prizes in 1978 and in 1997. Cited on page 78.

56 A picture of this unique clock can be found in the article by A. Garrett, Perpetual motion – a delicious delirium, *Physics World* pp. 23–26, December 1990. Cited on page 78.

57 A Shell study estimates the world's total energy consumption in 2000 to be 500 EJ. The US Department of Energy estimates it to be around 416 EJ. We took the lower value here. A discussion and a breakdown into electricity usage (14 EJ) and other energy forms, with variations per country, can be found in S. Benka, The energy challenge, *Physics Today* **55**, pp. 38–39, April 2002, and in E.J. Monitz & M.A. Kenderdine, Meeting energy challenges: technology and policy, *Physics Today* **55**, pp. 40–46, April 2002. Cited on page 80.

58 For an overview, see the paper by J.F. Mulligan & H.G. Hertz, An unpublished lecture by Heinrich Hertz: 'On the energy balance of the Earth', *American Journal of Physics* **65**, pp. 36–45, 1997. Cited on page 82.

59 For a beautiful photograph of this feline feat, see the cover of the journal and the article of J. Darius, A tale of a falling cat, *Nature* **308**, p. 109, 1984. Cited on page 86.

60 Natthi L. Sharma, A new observation about rolling motion, *European Journal of Physics* **17**, pp. 353–356, 1996. Cited on page 87.

61 C. Singh, When physical intuition fails, *American Journal of Physics* **70**, pp. 1103–1109, 2002. Cited on page 87.

62 Serge Gracovetsky, *The Spinal Engine*, Springer Verlag, 1990. Cited on page 88.

63 Thomas Heath, *Aristarchus of Samos – the Ancient Copernicus*, Dover, 1981, reprinted from the original 1913 edition. Aristarchos' treaty is given in Greek and English. Aristarchos was the first proposer of the heliocentric system. Aristarchos had measured the length of the day (in fact, by determining the number of days per year) to the astonishing precision of less than one second. This excellent book also gives an overview of Greek astronomy before Aristarchos, explained in detail for each Greek thinker. Aristarchos' text is also reprinted in Aristarchos, On the sizes and the distances of the sun and the moon, *c.* 280 BCE in Michael J. Crowe, Theories of the world from antiquity to the Copernican revolution, Dover, 1990, especially on pp. 27–29. Cited on pages 89 and 251.

64 The influence of the Coriolis effect on icebergs was studied most thoroughly by the Swedish physicist turned oceanographer Walfrid Ekman (1874–1954); the topic was suggested by the great explorer Fridtjof Nansen, who also made the first observations. In his honour, one speaks of Ekman layer, Ekman transport and Ekman spirals. Any text on oceanography or physical geography will tell more about them. Cited on page 91.

65 An overview of the effects of the Coriolis acceleration $\mathbf{a} = -2\omega \times \mathbf{v}$ in the rotating frame is given by Edward A. Desloge, *Classical Mechanics*, Volume 1, John Wiley & Sons, 1982. Even the *Gulf Stream*, the current of warm water flowing from the Caribbean to the North Sea, is influenced by it. Cited on page 91.

66 The original publication is by A.H. Shapiro, Bath-tub vortex, *Nature* **196**, pp. 1080–1081, 1962. He also produced two movies of the experiment. The experiment has been repeated many times in the northern and in the southern hemisphere, where the water drains clockwise; the first southern hemisphere test was L.M. Trefethen & al., The bath-tub vortex in the southern hemisphere, *Nature* **201**, pp. 1084–1085, 1965. A complete literature list is found in the letters to the editor of the *American Journal of Physics* **62**, p. 1063, 1994. Cited on page 91.

67 The tricks are explained by H. Richard Crane, Short Foucault pendulum: a way to eliminate the precession due to ellipticity, *American Journal of Physics* **49**, pp. 1004–1006, 1981, and particularly in H. Richard Crane, Foucault pendulum wall clock, *American Journal of Physics* **63**, pp. 33–39, 1993. The Foucault pendulum was also the topic of the thesis of Heike Kamerling Onnes, Nieuwe bewijzen der aswenteling der aarde, Universiteit Groningen, 1879. Cited on page 91.

68 The reference is J.G. Hagen, La rotation de la terre : ses preuves mécaniques anciennes et nouvelles, *Sp. Astr. Vaticana Second. App.* Rome, 1910. His other experiment is published as J.G. Hagen, How Atwoods machine shows the rotation of the Earth even quantitatively, International Congress of Mathematics, Aug. 1912. Cited on page 92.

69 R. Anderson, H.R. Bilger & G.E. Stedman, The Sagnac-effect: a century of Earth-rotated interferometers, *American Journal of Physics* **62**, pp. 975–985, 1994.

See also the clear and extensive paper by G.E. Stedman, Ring laser tests of fundamental physics and geophysics, *Reports on progress of physics* **60**, pp. 615–688, 1997. Cited on page 93.

70 About the length of the day, see the http://maia.usno.navy.mil website, or the books by K. Lambeck, *The Earth's Variable Rotation: Geophysical Causes and Consequences*, Cambridge University Press, 1980, and by W.H. Munk & G.J.F. MacDonald, *The Rotation of the Earth*, Cambridge University Press, 1960. Cited on pages 93 and 122.

71 One example of data is by C.P. Sonett, E.P. Kvale, A. Zakharian, M.A. Chan & T.M. Demko, Late proterozoic and paleozoic tides, retreat of the moon, and rotation of the Earth, *Science* **273**, pp. 100–104, 5 July 1996. They deduce from tidal sediment analysis that days were only 18 to 19 hours long in the Proterozoic, i.e. 900 million years ago; they assume that the year was 31 million seconds long all along. Another determination was by G.E. Williams, Precambrian tidal and glacial clastic deposits: implications for precambrian Earth–Moon dynamics and palaeoclimate, *Sedimentary Geology* **120**, pp. 55–74, 1998. Using a geological formation called *tidal rhythmites*, he deduced that about 600 million years ago there were 13 months per year and days had 22 hours. Cited on page 94.

72 The story of this combination of history and astronomy is told in Richard Stephenson, *Historical Eclispes and Earth's Rotation*, Cambridge University Press, 1996. Cited on page 94.

73 On the rotation and history of the solar system, see S. Brush, Theories of the origin of the solar system 1956–1985, *Reviews of Modern Physics* **62**, pp. 43–112, 1990. Cited on page 94.

74 The website http://maia.usno.navy.mil shows motion of the Earth's axis during the last ten years. The International Latitude Service founded by Küstner is now part of the International Earth Rotation Service; more information can be found on the http://www.iers.org website. The latest idea is that two-thirds of the circular component of the polar motion, which in the usa is called 'Chandler wobble' after the person who attributed to himself the discovery by Küstner, is due to fluctuations of the ocean pressure at the bottom of the oceans and one-third is due to pressure changes in the atmosphere of the Earth. This is explained by R.S. Gross, The excitation of the Chandler wobble, *Geophysical Physics Letters* **27**, pp. 2329–2332, 1st August 2000. Cited on page 95.

75 For more information about Alfred Wegener, see the (simple) text by Klaus Rohrbach, *Alfred Wegener – Erforscher der wandernden Kontinente*, Verlag Freies Geistesleben, 1993; about plate tectonics, see the http://www.scotese.com web site. About earthquakes, see the http://www.geo.ed.ac.uk/quakexe/quakes and the http://www.iris.edu/seismon website. See the http://vulcan.wr.usgs.gov and the http://www.dartmouth.edu/~volcano/ websites for information about volcanoes. Cited on page 96.

76 J.D. Hays, J. Imbrie & N.J. Shackleton, Variations in the Earth's orbit: pacemaker of the ice ages, *Science* **194**, pp. 1121–1132, 1976. They found the result by literally digging in the mud that covers the ocean floor in certain places. Note that the web is full of information on the ice ages. Just look up 'Milankovitch' in a search engine. Cited on page 98.

77 Roberta Humphreys & Jeffrey Larsen, The sun's distance above the galactic plane, *Astronomical Journal* **110**, pp. 2183–2188, November 1995. Cited on page 99.

78 C.L. Bennet, M.S. Turner & M. White, The cosmic rosetta stone, *Physics Today* **50**, pp. 32–38, November 1997. Cited on page 99.

79 A good roulette prediction story from the 1970s is told by Thomas A. Bass, *The Eudaemonic Pie* also published under the title *The Newtonian Casino*, Backinprint, 2000. An overview up to 1998 is given in the paper Edward O. Thorp, *The invention of the first wearable computer*, Proceeding of the Second International Symposium on Wearable Computers (ISWC 1998), 19-20 October 1998, Pittsburgh, Pennsylvania, USA (IEEE Computer Society), pp. 4–8, 1998, downloadable at http://csdl.computer.org/comp/proceedings/iswc/1998/9074/00/9074toc.htm. Cited on pages 101 and 751.

80 The original papers are A.H. Compton, A laboratory method of demonstrating the Earth's rotation, *Science* **37**, pp. 803–806, 1913, A.H. Compton, Watching the Earth revolve, *Scientific American Supplement* no. 2047, pp. 196–197, 1915, and A.H. Compton, A determination of latitude, azimuth and the length of the day independent of astronomical observations, *Physical Review* (second series) **5**, pp. 109–117, 1915. Cited on page 92.

81 This and many other physics surprises are described in the beautiful lecture script by Josef Zweck, *Physik im Alltag*, the notes of his lectures held in 1999/2000 at the Universität Regensburg. Cited on pages 101 and 104.

82 The equilibrium of ships, so important in car ferries, is an interesting part of shipbuilding; an introduction was already given by Leonhard Euler, Scientia navalis, 1749. Cited on page 102.

83 K.R. Weninger, B.P. Barber & S.J. Putterman, Pulsed Mie scattering measurements of the collapse of a sonoluminescing bubble, *Physical Review Letters* **78**, pp. 1799–1802, 1997. Cited on page 103.

84 On http://www.sff.net/people/geoffrey.landis/vacuum.html you can read a description of what happened. See also the http://www.sff.net/people/geoffrey.landis/ebullism.html and http://imagine.gsfc.nasa.gov/docs/ask_astro/answers/970603.html web sites. They all give details on the effects of vacuum on humans. Cited on page 103.

85 R. McN. Alexander, Leg design and jumping technique for humans, other vertebrates and insects, *Philosophical Transactions of the Royal Society in London B* **347**, pp. 235–249, 1995. Cited on page 107.

86 J. W. Glasheen & T. A. McMahon, A hydrodynamic model of locomotion in the basilisk lizard, *Nature* **380**, pp. 340–342, For pictures, see also *New Scientist*, p. 18, 30 March 1996, or *Scientific American*, pp. 48–49, September 1997, or the website by the author at http://rjf2.biol.berkeley.edu/Full_Lab/FL_Personnel/J_Glasheen/J_Glasheen.html.

Several shore birds also have the ability to run over water, using the same mechanism. Cited on page 107.

87 A. Fernandez–Nieves & F.J. de las Nieves, About the propulsion system of a kayak and of Basiliscus basiliscus, *European Journal of Physics* **19**, pp. 425–429, 1998. Cited on page 107.

88 The material on the shadow discussion is from the book by Robert M. Pryce, *Cook and Peary*, Stackpole Books, 1997. See also Wally Herbert, *The Noose of Laurels*, Doubleday 1989. The sad story of Robert Peary is also told in the centenary number of National Geographic, September 1988. Since the National Geographic Society had financed Peary in his attempt and had supported him until the US Congress had declared him the first man at the pole, the (partial) retraction is noteworthy. (The magazine then changed its mind again later on, to sell more copies.) By the way, the photographs of Cook, who claimed to have been at the North Pole even before Peary, have the same problem with the shadow length. Both men have a history of cheating about their 'exploits'. As a result, the first man at the North Pole was probably Roald Amundsen, who arrived there a few years later, and who was also the first man at the South Pole. Cited on page 108.

89 The story is told in M. Nauenberg, Hooke, orbital motion, and Newton's Principia, *American Journal of Physics* **62**, 1994, pp. 331–350. Cited on page 109.

90 More details are given by D. Rawlins, in Doubling your sunsets or how anyone can measure the Earth's size with wristwatch and meter stick, *American Journal of Physics* **47**, 1979, pp. 126–128. Another simple measurement of the Earth radius, using only a sextant, is given by R. O'Keefe & B. Ghavimi–Alagha, in The world trade centre and the distance to the world's centre, *American Journal of Physics* **60**, pp. 183–185, 1992. Cited on page 110.

91 More details on astronomical distance measurements can be found in the beautiful little book by A. van Helden, *Measuring the Universe*, University of Chicago Press, 1985, and in Nigel Henbest & Heather Cooper, *The Guide to the Galaxy*, Cambridge University Press, 1994. Cited on page 110.

92 A lot of details can be found in M. Jammer, *Concepts of Mass in Classical and Modern Physics*, reprinted by Dover, 1997, and in *Concepts of Force, a Study in the Foundations of Mechanics*, Harvard University Press, 1957. These eclectic and thoroughly researched texts provide numerous details and explain various philosophical viewpoints, but lack clear statements and conclusions on the accurate description of nature; thus are not of help on fundamental issues.

Jean Buridan (*c.* 1295 to *c.* 1366) criticizes the distinction of sublunar and translunar motion in his book *De Caelo*, one of his numerous works. Cited on page 110.

93 D. Topper & D.E. Vincent, An analysis of Newton's projectile diagram, *European Journal of Physics* **20**, pp. 59–66, 1999. Cited on page 110.

94 The absurd story of the metre is told in the historical novel by Ken Alder, *The Measure of All Things : The Seven-Year Odyssey and Hidden Error That Transformed the World*, The Free Press, 2003. Cited on page 112.

95 This is explained for example by D.K. Firpić & I.V. Aniçin, The planets, after all, may run only in perfect circles – but in the velocity space!, *European Journal of Physics* **14**, pp. 255–258, 1993. Cited on pages 127 and 1145.

96 About the measurement of spatial dimensions via gravity – and the failure to find any hint for a number different than three – see the review by E.G. Adelberger, B.R. Heckel & A.E. Nelson, Tests of the gravitational inverse-square law, *Annual Review of Nuclear and Particle Science* **53**, pp. 77–121, 2003, also http://www.arxiv.org/abs/hep-ph/0307284, or the review by J.A. Hewett & M. Spiropulu, Particle physics probes of extra spacetime dimensions, *Annual Review of Nuclear and Particle Science* **52**, pp. 397–424, 2002, http://www.arxiv.org/abs/hep-ph/0205106. Cited on page 114.

97 There are many books explaining the origin of the precise shape of the Earth, such as the pocket book S. Anders, *Weil die Erde rotiert*, Verlag Harri Deutsch, 1985. Cited on page 114.

98 The shape of the Earth is described most precisely with the World Geodetic System. For an extensive presentation, see the http://www.eurocontrol.be/projects/eatchip/wgs84/start.html web site. See also the website of the *International Earth Rotation Service* at http://hpiers.obspm.fr. Cited on page 114.

99 W.K. Hartman, R.J. Phillips & G.J. Taylor, editors, *Origin of the Moon*, Lunar and Planetary Institute, 1986. Cited on page 117.

100 Dietrich Neumann, Physiologische Uhren von Insekten – Zur Ökophysiologie lunarperiodisch kontrollierter Fortpflanzungszeiten, *Naturwissenschaften* **82**, pp. 310–320, 1995. Cited on page 117.

101 The origin of the duration of the menstrual cycle is not yet settled; however, there are explanations on how it gets synchronized with other cycles. For a general explanation see Arkady Pikovsky, Michael Rosenblum & Jürgen Kurths, *Synchronization: a Universal Concept in Nonlinear Science*, Cambridge University Press, 2002. Cited on page 118.

102 J. Laskkar, F. Joutel & P. Robutel, Stability of the Earth's obliquity by the moon, *Nature* **361**, pp. 615–617, 1993. However, the question is not completely settled, and other opinions exist. Cited on page 118.

103 Neil F. Comins, *What if the Moon Did not Exist? – Voyages to Earths that Might Have Been*, Harper Collins, 1993. Cited on page 118.

104 See for example the discussion by J.J. Lissauer, It is not easy to make the moon, *Nature* **389**, pp. 327–328, 1997. Cited on page 118.

105 Paul A. Wiegert, Kimmo A. Innanen & Seppo Mikkola, An asteroidal companion to the Earth, *Nature* **387**, pp. 685–686, 12 June 1997, together with the comment on pp. 651–652. Cited on page 120.

106 Simon Newcomb, *Astronomical Papers of the American Ephemeris* **1**, p. 472, 1882. Cited on page 120.

107 A beautiful introduction is the classic G. Falk & W. Ruppel, Mechanik, Relativität, Gravitation – ein Lehrbuch, Springer Verlag, Dritte Auflage, 1983. Cited on page 121.

108 J. Soldner, *Berliner Astronomisches Jahrbuch auf das Jahr 1804*, 1801, p. 161. Cited on page 123.

109 The equality was first tested with precision by Roland von Eőtvős, *Annalen der Physik & Chemie* **59**, p. 354, 1896, and by R. von Eőtvős, V. Pekár & E. Fekete, Beiträge zum Gesetz der Proportionalität von Trägheit und Gravität, *Annalen der Physik 4, Leipzig* **68**, pp. 11–66, 1922. He found agreement to 5 parts in 10^9. More experiments were performed by P.G. Roll, R. Krotkow & R.H. Dicke, The equivalence of inertial and passive gravitational mass, *Annals of Physics (NY)* **26**, pp. 442–517, 1964, one of the most interesting and entertaining research articles in experimental physics, and by V.B. Braginsky & V.I. Panov, *Soviet Physics – JETP* **34**, pp. 463–466, 1971. Modern results, with errors less than one part in 10^{12}, are by Y. Su & al., New tests of the universality of free fall, *Physical Review* **D50**, pp. 3614–3636, 1994. Several experiments have been proposed to test the equality in space to less than one part in 10^{16}. Cited on page 124.

110 See L. Hodges, Gravitational field strength inside the Earth, *American Journal of Physics* **59**, pp. 954–956, 1991. Cited on page 127.

111 P. Mohazzabi & M.C. James, Plumb line and the shape of the Earth, *American Journal of Physics* **68**, pp. 1038–1041, 2000. Cited on page 128.

112 From Neil de Gasse Tyson, *The Universe Down to Earth*, Columbia University Press, 1994. Cited on page 129.

113 G.D. Quinlan, Planet X: a myth exposed, *Nature* **363**, pp. 18–19, 1993. Cited on page 129.

114 See Robert Matthews, Not a snowball's chance ..., *New Scientist* 12 July 1997, pp. 24–27. The original claim is by Louis A. Frank, J.B. Sigwarth & J.D. Craven, On the influx of small comets into the Earth's upper atmosphere, parts I and II, *Geophysical Research Letters* **13**, pp. 303–306, pp. 307–310, 1986. The latest observations have disproved the claim. Cited on page 129.

115 The ray form is beautifully explained by J. Evans, The ray form of Newton's law of motion, *American Journal of Physics* **61**, pp. 347–350, 1993. Cited on page 130.

116 G.-L. Lesage, Lucrèce Newtonien, *Nouveaux mémoires de l'Académie Royale des Sciences et Belles Lettres* pp. 404–431, 1747. Cited on page 130.

117 G.J. Sussman & J. Wisdom, Chaotic Evolution of the Solar System, *Science* **257**, pp. 56–62, 1992. Cited on page 131.

118 Cassius Dio, *Historia Romana*, *c.* 220, book 37, 18. For an English translation, see the site http://penelope.uchicago.edu/Thayer/E/Roman/Texts/Cassius_Dio/37*.html. Cited on page 131.

119 D. Hestenes, M. Wells & G. Swackhamer, Force concept inventory, *Physics Teacher* **30**, pp. 141–158, 1982. The authors developed tests to check the understanding of the concept of physical force in students; the work has attracted a lot of attention in the field of physics teaching. Cited on page 135.

120 See M. Hirano, K. Shinjo, R. Kanecko & Y. Murata, Observation of superlubricity by scanning tunneling microscopy, *Physical Review Letters* **78**, pp. 1448–1451, 1997. See also the discussion of their results by Serge Fayeulle, Superlubricity: when friction stops, *Physics World* pp. 29–30, May 1997. Cited on page 136.

121 C. Donald Ahrens, *Meteorology Today: an Introduction to the Weather, Climate, and the Environment*, West Publishing Company, 1991. Cited on page 138.

122 This topic is discussed with lucidity by J.R. Mureika, What really are the best 100 m performances?, *Athletics: Canada's National Track and Field Running Magazine*, July 1997. It can also be found as http://www.arxiv.org/abs/physics/9705004, together with other papers on similar topics by the same author. Cited on page 138.

123 F.P. Bowden & D. Tabor, *The Friction and Lubrication of Solids*, Oxford University Press, Part I, revised edition, 1954, and part II, 1964. Cited on page 138.

124 The main tests of randomness of number series – among them the gorilla test – can be found in the authoritative paper by G. Marsaglia & W.W. Tsang, Some difficult-to-pass tests of randomness, *Journal of Statistical Software* 7, p. 8, 2002. It can also be downloaded from the http://www.jstatsoft.org/v07/i03/tuftests.pdf website. Cited on page 142.

125 For one aspect of the issue, see for example the captivating book by Bert Hellinger, *Zweierlei Glück*, Carl Auer Systeme Verlag, 1997. The author explains how to live serenely and with the highest possible responsibility of one's actions, by reducing entanglements with the destiny of others. He describes a simple and powerful technique to realise this goal.

A completely different viewpoint is given by Nobel Peace Price winner Aung San Suu Kyi, *Freedom from Fear*, Penguin, 1991. Cited on page 143.

126 Henrik Walter, *Neurophilosophie der Willensfreiheit*, Mentis Verlag, Paderborn 1999. Also available in English translation. Cited on page 144.

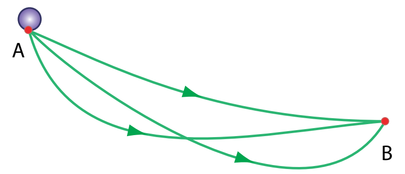


Figure 82 What shape of rail allows the black stone to glide most rapidly from point A to the lower point B?

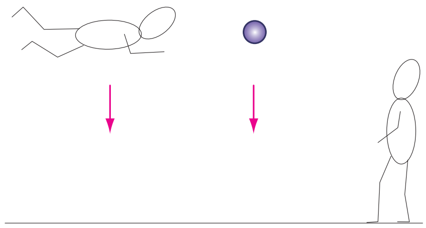


Figure 83 Can motion be described in a manner common to all observers?

4. Global descriptions of motion: the simplicity of complexity

> Πλεῖν ἀνάγκε, ζῆν οὐκ ἀνάγκη.*
>
> Pompeius

All over the Earth – even in Australia – people observe that stones fall 'down'. This ancient observation led to the discovery of the universal law of gravity. It was necessary to look for a description of gravity that was valid globally. The only additional observation that needs to be recognized in order to deduce the result $a = GM/r^2$ is the variation of gravity with height.

In short, thinking *globally* helps us to make our description of motion more precise. How can we describe motion as globally as possible?

We will describe six approaches to this question, each of which will be helpful on our way to the top of Motion Mountain. We will start with an overview, and then explore the details of each approach.

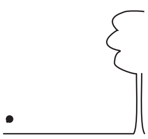
▪ The first global approach to motion arises from a limitation of what we have learned so far. When we predict the motion of a particle from its current acceleration, we are using the most *local* description of motion possible. For example, whenever we use an evolution equation we use the acceleration of a particle at a certain place and time to determine its position and motion *just after* that moment and *in the immediate neighbourhood* of that place.

Evolution equations thus have a mental 'horizon' of radius zero.

The opposite approach is illustrated in the famous problem of Figure 82. The challenge is to find the path that allows the fastest possible gliding motion from a high point to a
Challenge 319 ny distant low point. To solve this we need to consider the motion as a whole, for all times and positions. The global approach required by questions such as this one will lead us to a description of motion which is simple, precise and fascinating: the so-called principle of cosmic laziness, also known as the principle of least action.

▪ The second global approach to motion emerges when we compare the various descriptions of the same system produced by different observers. For example, the observations by somebody falling from a cliff, a passenger in a roller coaster, and an observer on

* Navigare necesse, vivere non necesse. 'To navigate is necessary, to live is not.' Gnaeus Pompeius Magnus
Ref. 54 (106–48 bce), as cited by Plutarchus (*c.* 45 to *c.* 125).



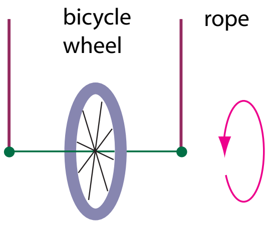


Figure 84 What happens when one rope is cut?

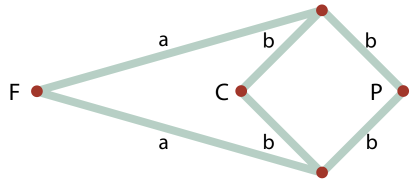


Figure 85 How to draw a straight line with a compass: fix point F, put a pencil into joint P and move C with a compass along a circle

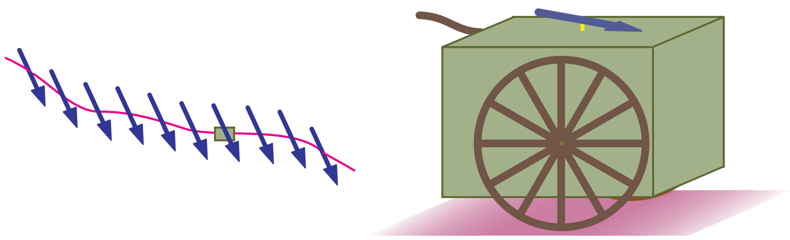


Figure 86 A south-pointing carriage

the ground will usually differ. The relationships between these observations lead us to a global description, valid for everybody. This approach leads us to the theory of relativity.

▪ The third global approach to motion is to exploring the motion of *extended and rigid* bodies, rather than mass points. The counter-intuitive result of the experiment in Figure 84 shows why this is worthwhile.

In order to design machines, it is essential to understand how a group of rigid bodies interact with one another. As an example, the mechanism in Figure 85 connects the motion of points C and P. It implicitly defines a circle such that one always has the relation Challenge 320 ny $r_C = 1/r_P$ between the distances of C and P from its centre. Can you find that circle?

Ref. 127 Another famous challenge is to devise a wooden carriage, with gearwheels that connect the wheels to an arrow in such a way that whatever path the carriage takes, the arrow Challenge 321 d always points south (see Figure 86). The solution to this is useful in helping us to understand general relativity, as we will see.

Another interesting example of rigid motion is the way that human movements, such Ref. 128 as the general motions of an arm, are composed from a small number of basic motions. All these examples are from the fascinating field of engineering; unfortunately, we will have little time to explore this topic in our hike.

▪ The fourth global approach to motion is the description of *non-rigid extended bodies*. For example, *fluid mechanics* studies the flow of fluids (like honey, water or air) around solid bodies (like spoons, ships, sails or wings). Fluid mechanics thus seeks to explain how Ref. 129 insects, birds and aeroplanes fly,* why sailboats can sail against the wind, what happens

* The mechanisms of insect flight are still a subject of active research. Traditionally, fluid dynamics has

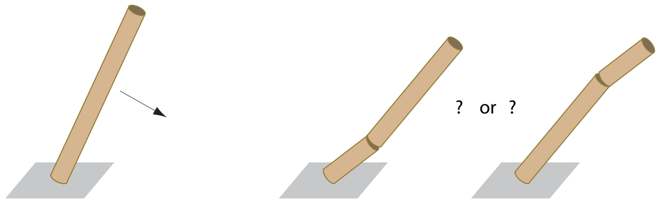


Figure 87 How and where does a falling brick chimney break?

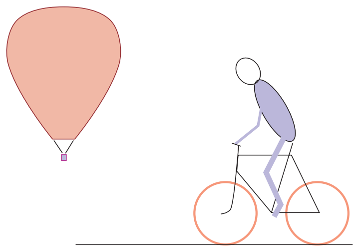


Figure 88 Why do hot-air balloons stay inflated? How can you measure the weight of a bicycle rider using only a ruler?

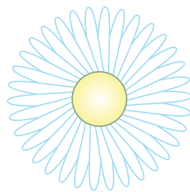


Figure 89 What determines the number of petals in a daisy?

when a hard-boiled egg is made to spin on a thin layer of water, or how a bottle full of wine can be emptied in the fastest way possible. Challenge 322 n

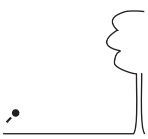
As well as fluids, we can study the behaviour of deformable *solids*. This area of research is called *continuum mechanics*. It deals with deformations and oscillations of extended structures. It seeks to explain, for example, why bells are made in particular shapes; how large bodies – such as falling chimneys – break when under stress; and how cats can turn themselves the right way up as they fall. Challenge 323 n During the course of our journey we will repeatedly encounter issues from this field, which impinges even upon general relativity and the world of elementary particles.

▪ The fifth global approach to motion is the study of the motion of huge numbers of particles. This is called *statistical mechanics*. The concepts needed to describe gases, such as temperature and pressure (see Figure 88), will be our first steps towards the understanding of black holes.

▪ The sixth global approach to motion involves all of the above-mentioned viewpoints *at the same time*. Such an approach is needed to understand everyday experience, and *life* itself. Why does a flower form a specific number of petals? How does an embryo differentiate in the womb What makes our hearts beat? How do mountains ridges and cloud patterns emerge? How do stars and galaxies evolve?

---

concentrated on large systems, like boats, ships and aeroplanes. Indeed, the smallest human-made object that can fly in a controlled way – say, a radio-controlled plane or helicopter – is much larger and heavier than many flying objects that evolution has engineered. It turns out that controlling the flight of small things requires more knowledge and more tricks than controlling the flight of large things. There is more about this topic on page 896.



How are sea waves formed by the wind?

All these are examples of *self-organization*; life scientists simply speak of *growth*. Whatever we call these processes, they are characterized by the spontaneous appearance of patterns, shapes and cycles. Such processes are a common research theme across many disciplines, including biology, chemistry medicine, geology and engineering.

We will now give a short introduction to these six global approaches to motion. We will begin with the first approach, namely, the global description of moving point-like objects. The beautiful method described below was the result of several centuries of collective effort, and is the highlight of mechanics. It also provides the basis for all the further descriptions of motion that we will meet later on.

Measuring change with action

Motion can be described by numbers. For a single particle, the relations between the spatial and temporal coordinates describe the motion. The realization that expressions like $(x(t), y(t), z(t))$ could be used to describe the path of a moving particle was a milestone in the development of modern physics.

We can go further. Motion is a type of change. And this change can itself be usefully described by numbers. In fact, change can be measured by a single number. This realization was the next important milestone. Physicists took almost two centuries of attempts to uncover the way to describe change. As a result, the quantity that measures change has a strange name: it is called *(physical) action*.* To remember the connection of 'action' with change, just think about a Hollywood movie: a lot of action means a large amount of change.

Imagine taking two snapshots of a system at different times. How could you define the amount of change that occurred in between? When do things change a lot, and when do they change only a little? First of all, a system with a lot of motion shows a lot of change. So it makes sense that the action of a system composed of independent subsystems should be the sum of the actions of these subsystems.

Secondly, change often – but not always – builds up over time; in other cases, recent change can compensate for previous change. Change can thus increase or decrease with time.

Thirdly, for a system in which motion is stored, transformed or shifted from one subsystem to another, the change is smaller than for a system where this is not the case.

* Note that this 'action' is not the same as the 'action' appearing in statements such as 'every action has an equal and opposite reaction'. This last usage, coined by Newton, has not stuck; therefore the term has been recycled. After Newton, the term 'action' was first used with an intermediate meaning, before it was finally given the modern meaning used here. This last meaning is the only meaning used in this text.

Another term that has been recycled is the 'principle of least action'. In old books it used to have a different meaning from the one in this chapter. Nowadays, it refers to what used to be called *Hamilton's principle* in the Anglo-Saxon world, even though it is (mostly) due to others, especially Leibniz. The old names and meanings are falling into disuse and are not continued here.

Behind these shifts in terminology is the story of an intense two-centuries-long attempt to describe motion with so-called *extremal* or *variational principles*: the objective was to complete and improve the work initiated by Leibniz. These principles are only of historical interest today, because all are special cases
Ref. 130 of the principle of least action described here.

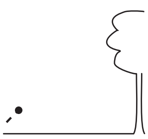


Table 20 Some action values for changes either observed or imagined

Change	Approximate action value
Smallest measurable change	$0.5 \cdot 10^{-34}$ Js
Exposure of photographic film	$1.1 \cdot 10^{-34}$ Js to 10^{-9} Js
Wing beat of a fruit fly	*c.* 1 pJs
Flower opening in the morning	*c.* 1 nJs
Getting a red face	*c.* 10 mJs
Held versus dropped glass	0.8 Js
Tree bent by the wind from one side to the other	500 Js
Making a white rabbit vanish by 'real' magic	100 PJs
Hiding a white rabbit	*c.* 0.1 Js
Maximum brain change in a minute	*c.* 5 Js
Levitating yourself within a minute by 1 m	*c.* 40 kJs
Car crash	*c.* 2 kJs
Birth	*c.* 2 kJs
Change due to a human life	*c.* 1 EJs
Driving car stops within the blink of an eye	20 kJs
Large earthquake	*c.* 1 PJs
Driving car disappears within the blink of an eye	1 ZJs
Sunrise	*c.* 0.1 ZJs
Gamma ray burster before and after explosion	*c.* 10^{46} Js
Universe after one second has elapsed	undefined and undefinable

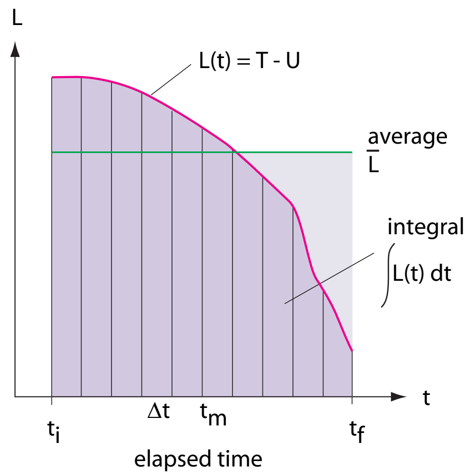
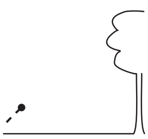


Figure 90 Defining a total effect as an accumulation (addition, or integral) of small effects over time

The mentioned properties imply that the natural measure of

Joseph Lagrange



change is the average difference between kinetic and potential energy multiplied by the elapsed time. This quantity has all the right properties: it is (usually) the sum of the corresponding quantities for all subsystems if these are independent; it generally increases with time (unless the evolution compensates for something that happened earlier); and it decreases if the system transforms motion into potential energy.

Challenge 324 e

Thus the (physical) *action* S, measuring the change in a system, is defined as

$$S = \overline{L} \cdot (t_\mathrm{f} - t_\mathrm{i}) = \overline{T - U} \cdot (t_\mathrm{f} - t_\mathrm{i}) = \int_{t_\mathrm{i}}^{t_\mathrm{f}} (T - U)\, \mathrm{d}t = \int_{t_\mathrm{i}}^{t_\mathrm{f}} L\, \mathrm{d}t\ , \tag{58}$$

Page 113

where T is the kinetic energy, U the potential energy we already know, L is the difference between these, and the overbar indicates a time average. The quantity L is called the *Lagrangian (function)* of the system,* describes what is being added over time, whenever things change. The sign $\int$ is a stretched 'S', for 'sum', and is pronounced 'integral of'. In intuitive terms it designates the operation (called *integration*) of adding up the values of a varying quantity in infinitesimal time steps $\mathrm{d}t$. The initial and the final times are written below and above the integration sign, respectively. Figure 90 illustrates the idea: the integral is simply the size of the dark area below the curve $L(t)$.

Challenge 325 e

Mathematically, the integral of the curve $L(t)$ is defined as

$$\int_{t_\mathrm{i}}^{t_\mathrm{f}} L(t)\, \mathrm{d}t = \lim_{\Delta t \to 0} \sum_{\mathrm{m=i}}^{\mathrm{f}} L(t_\mathrm{m}) \Delta t = \overline{L} \cdot (t_\mathrm{f} - t_\mathrm{i})\ . \tag{59}$$

In other words, the integral is the limit, as the time slices get smaller, of the sum of the areas of the individual rectangular strips that approximate the function.** Since the $\sum$ sign also means a sum, and since an infinitesimal Δt is written $\mathrm{d}t$, we can understand the notation used for integration. Integration is a sum over slices. The notation was developed by Gottfried Leibniz to make exactly this point. Physically speaking, the integral of the Lagrangian measures the *effect* that L builds up over time. Indeed, action is called 'effect' in some languages, such as German.

In short, then, action is the integral of the Lagrangian over time.

The unit of action, and thus of physical change, is the unit of energy (the Joule), times the unit of time (the second). Thus change is measured in Js. A large value means a big change. Table 20 shows some approximate values of actions.

To understand the definition of action in more detail, we will start with the simplest case: a system for which the potential energy is zero, such as a particle moving freely. Ob-

* It is named after Giuseppe Lodovico Lagrangia (b. 1736 Torino, d. 1813 Paris), better known as Joseph Louis Lagrange. He was the most important mathematician of his time; he started his career in Turin, then worked for 20 years in Berlin, and finally for 26 years in Paris. Among other things he worked on number theory and analytical mechanics, where he developed most of the mathematical tools used nowadays for calculations in classical mechanics and classical gravitation. He applied them successfully to many motions in the solar system.

** For more details on integration see Appendix D.

viously, a large kinetic energy means a lot of change. If we observe the particle at two instants, the more distant they are the larger the change. Furthermore, the observed change is larger if the particle moves more rapidly, as its kinetic energy is larger. This is not surprising.

Next, we explore a single particle moving in a potential. For example, a falling stone loses potential energy in exchange for a gain in kinetic energy. The more energy is exchanged, the more change there is. Hence the minus sign in the definition of L. If we explore a particle that is first thrown up in the air and then falls, the curve for $L(t)$ first is below the times axis, then above. We note that the definition of integration makes us count the grey surface *below* the time axis *negatively*. Change can thus be negative, and be compensated by subsequent change, as expected.

To measure change for a system made of several independent components, we simply add all the kinetic energies and subtract all the potential energies. This technique allows us to define actions for gases, liquids and solid matter. Even if the components interact, we still get a sensible result. In short, action is an *additive* quantity.

Physical action thus measures, in a single number, the change observed in a system between two instants of time. The observation may be anything at all: an explosion, a caress or a colour change. We will discover later that this idea is also applicable in relativity and quantum theory. Any change going on in any system of nature can be measured with a single number.

The principle of least action

We now have a precise measure of change, which, as it turns out, allows a simple and powerful description of motion. In nature, the change happening between two instants is always the *smallest* possible. *In nature, action is minimal.** Of all possible motions, nature always chooses for which the change is *minimal*. Let us study a few examples.

In the simple case of a free particle, when no potentials are involved, the principle of minimal action implies that the particle moves in a *straight* line with *constant* velocity. All other paths would lead to larger actions. Can you verify this?

Challenge 326 e

When gravity is present, a thrown stone flies along a parabola (or more precisely, along an ellipse) because any other path, say one in which the stone makes a loop in the air, would imply a *larger* action. Again you might want to verify this for yourself.

Challenge 327 e

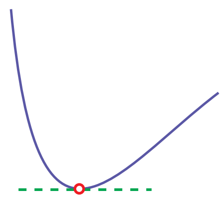


Figure 91 The minimum of a curve has vanishing slope

All observations support this simple and basic statement: things always move in a way that produces the smallest possible value for the action. This statement applies to the full path and to any of its segments. Betrand Russell called it the 'law of cosmic laziness'.

It is customary to express the idea of minimal change in a different way. The action varies when the path is varied. The actual path is the one with the smallest action. You will recall from school that at a minimum the derivative of a quantity vanishes: a minimum

* In fact, in some pathological situations the action is maximal, so that the snobbish form of the principle is that the action is 'stationary,' or an 'extremum,' meaning minimal *or* maximal. The condition of vanishing variation, given below, encompasses both cases.

has a horizontal slope. In the present case, we do not vary a quantity, but a complete path; hence we do not speak of a derivative or slope, but of a variation. It is customary to write the variation of action as δS. The *principle of least action* thus states:

$$\triangleright \textit{ The actual trajectory between specified end points satisfies } \delta S = 0 . \tag{60}$$

Mathematicians call this a *variational principle*. Note that the end points have to be specified: we have to compare motions with the *same* initial and final situations.

Before discussing the principle further, we can check that it is equivalent to the evolution equation.* To do this, we can use a standard procedure, part of the so-called *calculus*

---

* For those interested, here are a few comments on the equivalence of Lagrangians and evolution equations. Page 136 First of all, Lagrangians do not exist for non-conservative, or *dissipative* systems. We saw that there is no potential for any motion involving *friction* (and more than one dimension); therefore there is no action in these cases. One approach to overcome this limitation is to use a generalized formulation of the principle of least action. Whenever there is no potential, we can express the *work* variation δW between different trajectories x_i as

$$\delta W = \sum_i m_i \ddot{x}_i \delta x_i \; . \tag{61}$$

Motion is then described in the following way:

$$\triangleright \textit{ The actual trajectory satisfies } \int_{t_i}^{t_f} (\delta T + \delta W) \mathrm{d}t = 0 \quad \textit{provided} \quad \delta x(t_i) = \delta x(t_f) = 0 \; . \tag{62}$$

Challenge 328 ny The quantity being varied has no name; it represents a generalized notion of change. You might want to check that it leads to the correct evolution equations. Thus, although *proper* Lagrangian descriptions exist only for *conservative* systems, for dissipative systems the principle can be generalized and remains useful.

Many physicists will prefer another approach. What a mathematician calls a generalization is a special case for a physicist: the principle (62) hides the fact that *all* friction results from the usual principle of minimal action, if we include the complete microscopic details. There is no friction in the microscopic domain. Friction is an approximate, macroscopic concept.

Nevertheless, more mathematical viewpoints are useful. For example, they lead to interesting limitations for the use of Lagrangians. These limitations, which apply only if the world is viewed as purely classical – which it isn't – were discovered about a hundred years ago. In those times computers where not available, and the exploration of new calculation techniques was important. Here is a summary.

The coordinates used in connection with Lagrangians are not necessarily the Cartesian ones. *Generalized* coordinates are especially useful when there are *constraints* on the motion. This is the case for a pendulum, where the weight always has to be at the same distance from the suspension, or for an ice skater, where the Ref. 131 skate has to move in the direction in which it is pointing. Generalized coordinates may even be mixtures of positions and momenta. They can be divided into a few general types.

Generalized coordinates are called *holonomic–scleronomic* if they are related to Cartesian coordinates in a fixed way, independently of time: physical systems described by such coordinates include the pendulum and a particle in a potential. Coordinates are called *holonomic–rheonomic* if the dependence involves time. An example of a rheonomic systems would be a pendulum whose length depends on time. The two terms Page 221 rheonomic and scleronomic are due to Ludwig Boltzmann. These two cases, which concern systems that are only described by their geometry, are grouped together as *holonomic systems*. The term is due to Heinrich Page 517 Hertz.

The more general situation is called *anholonomic*, or *nonholonomic*. Lagrangians work well only for holonomic systems. Unfortunately, the meaning of the term 'nonholonomic' has changed. Nowadays, the term is also used for certain rheonomic systems. The modern use calls nonholonomic any system which involves velocities. Therefore, an ice skater or a rolling disk is often called a nonholonomic system. Care is thus necessary to decide what is meant by nonholonomic in any particular context.

Even though the use of Lagrangians, and of action, has its limitations, these need not bother us at micro-

of variations. The condition $\delta S = 0$ implies that the action, i.e. the area under the curve in Figure 90, is a minimum. A little bit of thinking shows that if the Lagrangian is of the form $L(x_n, v_n) = T(v_n) - U(x_n)$, then

Challenge 329 ny

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial v_n}\right) = \frac{\partial U}{\partial x_n} \tag{63}$$

where n counts all coordinates of all particles.* For a single particle, these *Lagrange's equations of motion* reduce to

Challenge 330 e

$$m\mathbf{a} = \nabla U \ . \tag{65}$$

This is the evolution equation: it says that the force on a particle is the gradient of the potential energy U. The principle of least action thus implies the equation of motion. (Can you show the converse?)

Challenge 331 n

In other words, *all systems evolve in such a way that the change is as small as possible.* Nature is economical. Nature is thus the opposite of a Hollywood thriller, in which the action is maximized; nature is more like a wise old man who keeps his actions to a minimum.

The principle of minimal action also states that the actual trajectory is the one for which the *average* of the Lagrangian over the whole trajectory is minimal (see Figure 90). Nature is a Dr. Dolittle. Can you verify this? This viewpoint allows one to deduce Lagrange's equations (63) directly.

Challenge 332 ny

The principle of least action distinguishes the actual trajectory from all other imaginable ones. This observation lead Leibniz to his famous interpretation that the actual world is the 'best of all possible worlds.'** We may dismiss this as metaphysical speculation, but we should still be able to feel the fascination of the issue. Leibniz was so excited about the principle of least action because it was the first time that actual observations were distinguished from all other imaginable possibilities. For the first time, the search for reasons why things are the way they are became a part of physical investigation. Could the world be different from what it is? In the principle of least action, we have a hint of a negative answer. (What do you think?) The final answer will emerge only in the last part of our adventure.

Challenge 333 n

As a way to describe motion, the Lagrangian has several advantages over the evolution equation. First of all, the Lagrangian is usually more *compact* than writing the corres-

scopic level, since microscopic systems are always conservative, holonomic and scleronomic. At the fundamental level, evolution equations and Lagrangians are indeed equivalent.

* The most general form for a Lagrangian $L(q_n, \dot{q}_n, t)$, using generalized holonomic coordinates q_n, leads to Lagrange equations of the form

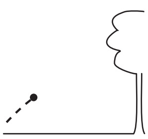
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_n}\right) = \frac{\partial L}{\partial q_n} \ . \tag{64}$$

In order to deduce these equations, we also need the relation $\delta\dot{q} = \mathrm{d}/\mathrm{d}t(\delta q)$. This relation is valid only for *holonomic* coordinates introduced in the previous footnote and explains their importance.

It should also be noted that the Lagrangian for a moving system is not unique; however, the study of how the various Lagrangians for a given moving system are related is not part of this walk.

Ref. 132

** This idea was ridiculed by the French philosopher Voltaire (1694–1778) in his lucid writings, notably in the brilliant book *Candide*, written in 1759, and still widely available.



ponding evolution equations. For example, only *one* Lagrangian is needed for one system, however many particles it includes. One makes fewer mistakes, especially sign mistakes, as one rapidly learns when performing calculations. Just try to write down the evolution equations for a chain of masses connected by springs; then compare the effort with a derivation using a Lagrangian. (The system behaves like a chain of atoms.) Challenge 334 ny We will encounter another example shortly: David Hilbert took only a few weeks to deduce the equations of motion of general relativity using a Lagrangian, whereas Albert Einstein had worked for ten years searching for them directly.

In addition, the description with a Lagrangian is valid with *any* set of coordinates describing the objects of investigation. The coordinates do not have to be Cartesian; they can be chosen as one prefers: cylindrical, spherical, hyperbolic, etc. These so-called *generalized coordinates* allow one to rapidly calculate the behaviour of many mechanical systems that are in practice too complicated to be described with Cartesian coordinates. For example, for programming the motion of robot arms, the angles of the joints provide a clearer description than Cartesian coordinates of the ends of the arms. Angles are non-Cartesian coordinates. They simplify calculations considerably: the task of finding the most economical way to move the hand of a robot from one point to another can be solved much more easily with angular variables.

More importantly, the Lagrangian allows one to quickly deduce the essential properties of a system, namely, its *symmetries* and its *conserved quantities*. We will develop this important idea shortly, and use it regularly throughout our walk. Page 179

Finally, the Lagrangian formulation can be generalized to encompass *all types of interactions*. Since the concepts of kinetic and potential energy are general, the principle of least action can be used in electricity, magnetism and optics as well as mechanics. The principle of least action is central to general relativity and to quantum theory, and allows one to easily relate both fields to classical mechanics.

As the principle of least action became well known, people applied it to an ever-increasing number of problems. Ref. 130 Today, Lagrangians are used in everything from the study of elementary particle collisions to the programming of robot motion in artificial intelligence. However, we should not forget that despite its remarkable simplicity and usefulness, the Lagrangian formulation is *equivalent* to the evolution equations. It is neither more general nor more specific. Challenge 335 n In particular, it is *not an explanation* for any type of motion, but only a view of it. In fact, the search of a new physical 'law' of motion is *just* the search for a new Lagrangian. This makes sense, as the description of nature always requires the description of change. Change in nature is always described by actions and Lagrangians.

The principle of least action states that the action is minimal when the end point of the motion, and in particular the time between them, are fixed. Ref. 133 It is less well known that the reciprocal principle also holds: if the action is kept fixed, the elapsed time is maximal. Challenge 336 ny Can you show this?

Even though the principle of least action is not an explanation of motion, it somehow calls for one. We need some patience, though. *Why* nature follows the principle of least action, and *how* it does so, will become clear when we explore quantum theory.

> Never confuse movement with action.
> Ernest Hemingway Ref. 136

Why is motion so often bounded?

> The optimist thinks this is the best of all possible worlds, and the pessimist knows it.
>
> Robert Oppenheimer

Looking around ourselves on Earth and in the sky, we find that matter is not evenly distributed. Matter tends to be near other matter: it is lumped together in *aggregates*. Some Ref. 137 major examples of aggregates are given in Figure 92 and Table 21. In the mass–size diagram of Figure 92, both scales are logarithmic. One notes three straight lines: a line $m \sim l$ extending from the Planck mass* upwards, via black holes, to the universe itself; a line $m \sim 1/l$ extending from the Planck mass downwards, to the lightest possible aggregate; and the usual matter line with $m \sim l^3$, extending from atoms upwards, via the Earth and the Sun. The first of the lines, the black hole limit, is explained by general relativity; the last two, the aggregate limit and the common matter line, by quantum theory.**

The aggregates outside the common matter line also show that the stronger the interaction that keeps the components together, the smaller the aggregate. But why is matter mainly found in lumps?

First of all, aggregates form because of the existence of *attractive* interactions between objects. Secondly, they form because of *friction*: when two components approach, an aggregate can only be formed if the released energy can be changed into heat. Thirdly, aggregates have a finite size because of *repulsive* effects that prevent the components from collapsing completely. Together, these three factors ensure that bound motion is much more common than unbound, 'free' motion.

Only three types of attraction lead to aggregates: gravity, the attraction of electric charges, and the strong nuclear interaction. Similarly, only three types of repulsion are observed: rotation, pressure, and the Pauli exclusion principle (which we will encounter Page 732 later on). Of the nine possible combinations of attraction and repulsion, not all appear in Challenge 337 n nature. Can you find out which ones are missing from Figure 92 and Table 21, and why?

Together, attraction, friction and repulsion imply that change and action are minimized when objects come and stay together. The principle of least action thus implies the stability of aggregates. By the way, formation history also explains why so many aggregates *rotate*. Can you tell why? Challenge 338 ny

But why does friction exist at all? And why do attractive and repulsive interactions exist? And why is it – as it would appear from the above – that in some distant past matter was *not* found in lumps? In order to answer these questions, we must first study another global property of motion: symmetry.

Table 21 Some major aggregates observed in nature

Aggregate	Size (diameter)	Obs. num.	Constituents
gravitationally bound aggregates			

Page 1063

* The Planck mass is given by $m_{\mathrm{Pl}} = \sqrt{\hbar c/G} = 21.767(16)\ \mu\mathrm{g}$.

** Figure 92 suggests that domains beyond physics exist; we will discover later on that this is not the case, as mass and size are not definable in those domains.

Aggregate	Size (diameter)	Obs. num.	Constituents
matter across universe	*c.* 100 Ym	1	superclusters of galaxies, hydrogen andhelium atoms
quasar	10^{12} to 10^{14} m	$20 \cdot 10^{6}$	baryons and leptons
supercluster of galaxies	*c.* 3 Ym	10^{7}	galaxy groups and clusters
galaxy cluster	*c.* 60 Zm	$25 \cdot 10^{9}$	10 to 50 galaxies
galaxy group or cluster	*c.* 240 Zm		50 to over 2000 galaxies
our local galaxy group	50 Zm	1	*c.* 40 galaxies
general galaxy	0.5 to 2 Zm	$3.5 \cdot 10^{12}$	10^{10} to $3 \cdot 10^{11}$ stars, dust and gas clouds, probably solar systems
our galaxy	1.0(0.1) Zm	1	10^{11} stars, dust and gas clouds, solar systems
interstellar clouds	up to 15 Am	$\gg 10^{5}$	hydrogen, ice and dust
solar system [a]	unknown	> 100	star, planets
our solar system	30 Pm	1	Sun, planets (Pluto's orbit's diameter: 11.8 Tm), moons, planetoids, comets, asteroids, dust, gas
Oort cloud	6 to 30 Pm	1	comets, dust
Kuiper belt	60 Tm	1	planetoids, comets, dust
star[b]	10 km to 100 Gm	$10^{22\pm1}$	ionized gas: protons, neutrons, electrons, neutrinos, photons
our star	1.39 Gm		
planet [a] (Jupiter, Earth)	143 Mm, 12.8 Mm	9+*c.* 100	solids, liquids, gases; in particular, heavy atoms
planetoids (Varuna, etc)	50 to 1 000 km	*c.* 10 (est. 10^{9})	solids
moons	10 to 1 000 km	*c.* 50	solids
neutron stars	10 km	*c.* 1000	mainly neutrons
electromagnetically bound aggregates [c]			
asteroids, mountains[d]	1 m to 930 km	>26 000	(10^{9} estimated) solids, usually monolithic
comets	10 cm to 50 km	> 10^{6}	ice and dust
planetoids, solids, liquids, gases, cheese	1 nm to > 100 km	n.a.	molecules, atoms
animals, plants, kefir	5 µm to 1 km	$10^{26\pm2}$	organs, cells
brain	0.15 m	10^{10}	neurons and other cell types
cells:		$10^{31\pm1}$	organelles, membranes, molecules
smallest (nanobacteria)	*c.* 5 µm		molecules
amoeba	600 µm		molecules
largest (whale nerve, single-celled plants)	*c.* 30 m		molecules
molecules:		*c.* $10^{78\pm2}$	atoms

Aggregate	Size (diameter)	Obs. num.	Constituents
H_2	*c.* 50 pm	$10^{72\pm2}$	atoms
DNA (human)	2 m (total per cell)	10^{21}	atoms
atoms, ions	30 pm to 300 pm	$10^{80\pm2}$	electrons and nuclei
aggregates bound by the weak interaction [c]			
none			
aggregates bound by the strong interaction [c]			
nucleus	$> 10^{-15}$ m	$10^{79\pm2}$	nucleons
nucleon (proton, neutron)	*c.* 10^{-15} m	$10^{80\pm2}$	quarks
mesons	*c.* 10^{-15} m	n.a.	quarks
neutron stars: see above			

a. Only in 1994 was the first evidence found for objects circling stars other than our Sun; of over 100 *extrasolar planets* found so far, most are found around F, G and K stars, including neutron stars. For example, three objects circle the pulsar PSR 1257+12, and a matter ring circles the star Ref. 138 β Pictoris. The objects seem to be dark stars, brown dwarfs or large gas planets like Jupiter. Due to the limitations of observation systems, none of the systems found so far form solar systems of the type we live in. In fact, only a few Earth-like planets have been found so far.

b. The Sun is among the brightest 7 % of stars. Of all stars, 80 %, are red M dwarfs, 8 % are orange K dwarfs, and 5 % are white D dwarfs: these are all faint. Almost all stars visible in the night sky belong to the bright 7 %. Some of these are from the rare blue O class or blue B class (such as Spica, Regulus and Riga); 0.7 % consist of the bright, white A class (such as Sirius, Vega and Altair); 2 % are of the yellow–white F class (such as Canopus, Procyon and Polaris); 3.5 % are of the yellow G class (like Alpha Centauri, Capella or the Sun). Exceptions include the few visible K giants, such Page 430 as Arcturus and Aldebaran, and the rare M supergiants, such as Betelgeuse and Antares. More on stars later on.

c. For more details on *microscopic* aggregates, see the table of composites in Appendix C.

d. It is estimated that there are about 10^9 asteroids (or planetoids) larger than 1 km and about Ref. 139 10^{20} that are heavier than 100 kg. By the way, no asteroids between Mercury and the Sun – the hypothetical *Vulcanoids* – have been found so far.

Curiosities and fun challenges about Lagrangians

Lagrangians and variational principles form a fascinating topic, which has charmed physicists for the last four centuries.

▪ When Lagrange published his book *Mécanique analytique*, in 1788, it formed one of the high points in the history of mechanics. He was proud of having written a systematic exposition of mechanics without a single figure. Obviously the book was difficult to read and was not a sales success. Therefore his methods took another generation to come into general use.

▪ Given that action is the basic quantity describing motion, we can define energy as action per unit time, and momentum as action per unit distance. The *energy* of a system

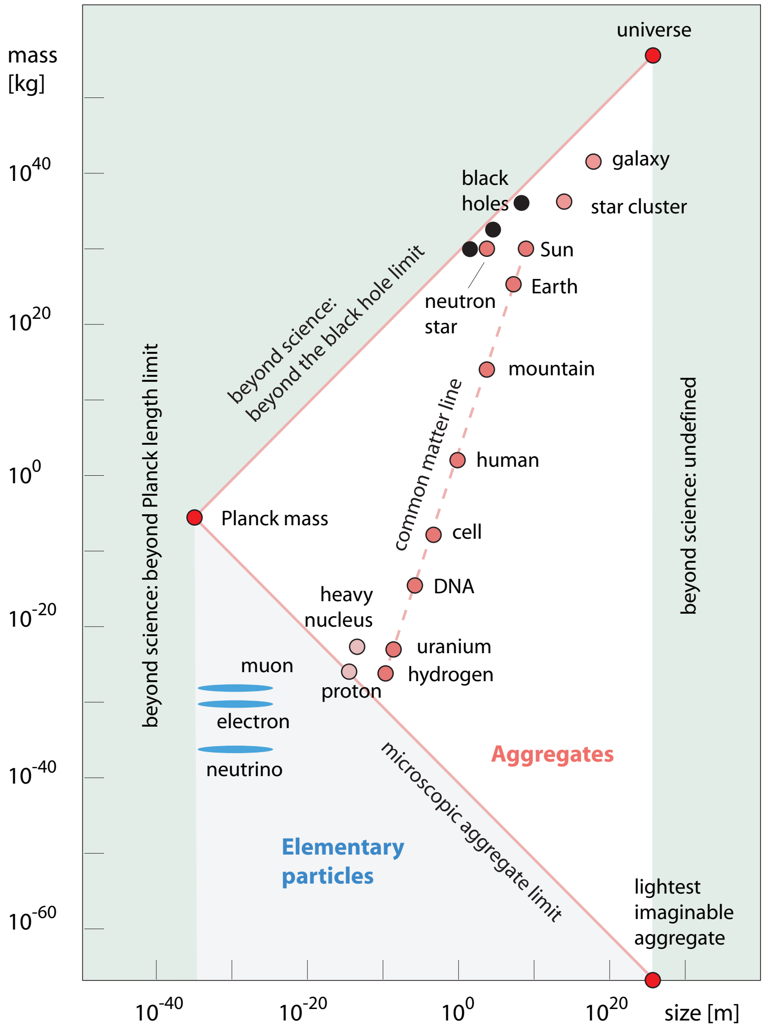


Figure 92 Aggregates in nature

thus describes how much it changes over time, and the *momentum* how much it changes over distance. What are angular momentum and rotational energy?

Challenge 339 n

▪ 'In nature, effects of telekinesis or prayer are impossible, as in most cases the change inside the brain is much smaller than the change claimed in the outside world.' Is this argument correct?

Challenge 340 n

▪ In Galilean physics, the Lagrangian is the difference between kinetic and potential energy. Later on, this definition will be generalized in a way that sharpens our understanding of this distinction: the Lagrangian becomes the difference between a term for free particles and a term due to their interactions. In other words, particle motion is a continuous compromise between what the particle would do if it were free and what other particles want it to do. In this respect, particles behave a lot like humans beings.

Challenge 341 ny

▪ Explain: why is $T + U$ constant, whereas $T - U$ is minimal?

▪ In nature, the sum $T + U$ of kinetic and potential energy is *constant* during motion (for closed systems), whereas the average of the difference $T - U$ is *minimal*. Is it possible to deduce, by combining these two facts, that systems tend to a state with minimum potential energy? Challenge 342 ny

▪ There is a principle of *least effort* describing the growth of trees. When a tree – a *monopodal phanerophyte* – grows and produces leaves, between 40% and 60% of the mass it consists of, namely the water and the minerals, has to be lifted upwards from the ground.* Therefore, a tree gets as many branches as high up in the air as possible using the smallest amount of energy. This is the reason why not all leaves are at the very top of a tree. Can you deduce more details about trees from this principle? Challenge 343 ny

▪ Another minimization principle can be used to understand the construction of animal bodies, especially their size and the proportions of their inner structures. (Ref. 134) For example, the heart pulse and breathing frequency both vary with animal mass m as $m^{-1/4}$, and the dissipated power varies as $m^{3/4}$. It turns out that such exponents result from three properties of living beings. First, they transport energy and material through the organism via a branched network of vessels: a few large ones, and increasingly many smaller ones. Secondly, the vessels all have the same minimum size. And thirdly, the networks are optimized in order to minimize the energy needed for transport. Together, these relations explain many additional scaling rules; they might also explain why animal lifespan scales as $m^{-1/4}$, or why most mammals have roughly the same number of heart beats in a lifetime.

A competing explanation, using a different minimization principle, states that quarter powers arise in any network built in order that the flow arrives to the destination by the most direct path. (Ref. 135)

▪ The minimization principle for the motion of light is even more beautiful: light always takes the path that requires the shortest travel time. It was known long ago that this idea describes exactly how light changes direction when it moves from air to water. In water, light moves more slowly; the speed ratio between air and water is called the *refractive index* of water. The refractive index, usually abbreviated n, is material-dependent. The value for water is about 1.3. This speed ratio, together with the minimum-time principle, leads to the 'law' of refraction, a simple relation between the sines of the two angles. Can you deduce it? (Challenge 344 n) (In fact, the exact definition of the refractive index is with respect to vacuum, not to air. But the difference is negligible: can you imagine why?) (Challenge 345 n)

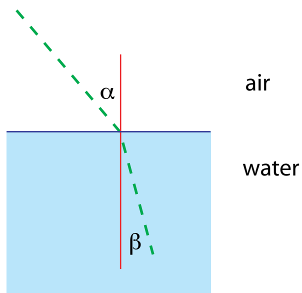


Figure 93 Refraction of light is due to travel-time optimization

For diamond, the refractive index is 2.4. The high value is one reason for the sparkle of diamonds cut with the 57-face *brilliant* cut. Can you think of some other reasons? Challenge 346 n

▪ Can you confirm that each of these minimization principles is a special case of the principle of least action? (Challenge 347 n) In fact, this is the case for *all* known minimization principles in nature. Each of them, like the principle of least action, is a principle of least change.

▪ In Galilean physics, the value of the action depends on the speed of the observer,

* The rest of the mass comes form the CO_2 in the air.

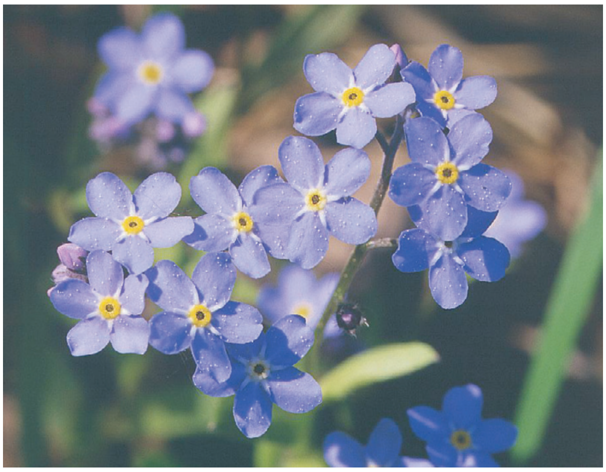


Figure 94 Forget-me-not, also called *Myosotis* (Boraginaceae)

but not on his position or orientation. But the action, when properly defined, should *not* depend on the observer. All observers should agree on the value of the observed change. Only special relativity will fulfil the requirement that action be independent of
Challenge 348 n the observer's speed. How will the relativistic action be defined?

▪ Measuring all the change that is going on in the universe presupposes that the uni-
Challenge 349 n verse is a physical system. Is this the case?

Motion and symmetry

The second way to describe motion globally is to describe it in such a way that *all* observers agree. An object under observation is called *symmetric* if it looks the same when seen from different points of view. For example, a forget-me-not flower, shown in Figure 94, is symmetrical because it looks the same after turning around it by 72 degrees; many fruit tree flowers have the same symmetry. One also says that under change of viewpoint the flower has an *invariant property*, namely its shape. If many such viewpoints are possible, one talks about a *high* symmetry, otherwise a *low* symmetry. For example, a four-leaf clover has a higher symmetry than a usual, three-leaf one. Different points of view imply different observers; in physics, the viewpoints are often called *frames of reference* and are described mathematically by coordinate systems.

High symmetry means many agreeing observers. At first sight, not many objects or observations in nature seem to be symmetrical. But this is a mistake. On the contrary, we can deduce that nature as a whole is symmetric from the simple fact that we have the
Challenge 350 n ability to talk about it! Moreover, the symmetry of nature is considerably higher than that of a forget-me-not. We will discover that this high symmetry is at the basis of the famous expression $E_0 = mc^2$.

Why can we think and talk?

> The hidden harmony is stronger than the apparent.
> Ref. 19 Heraclitos of Ephesos, about 500 BCE

Why can we understand somebody when he is talking about the world, even though we are not in his shoes? We can for two reasons: because most things look *similar* from different viewpoints, and because most of us have already had similar experiences *beforehand*.

'Similar' means that what *we* and what *others* observe somehow correspond. In other words, many aspects of observations do not depend on viewpoint. For example, the number of petals of a flower has the same value for all observers. We can therefore say that this quantity has the highest possible symmetry. We will see below that mass is another such example. Observables with the highest possible symmetry are called *scalars* in physics. Other aspects change from observer to observer. For example, the apparent size varies with the distance of observation. However, the actual size is observer-independent. In general terms, any type of *viewpoint-independence* is a form of symmetry, and the observation that two people looking at the same thing from different viewpoints can understand each other proves that nature is symmetric. We start to explore the details of this symmetry in this section and we will continue during most of the rest of our hike.

In the world around us, we note another general property: not only does the same phenomenon look similar to different observers, but *different* phenomena look similar to the *same* observer. For example, we know that if fire burns the finger in the kitchen, it will do so outside the house as well, and also in other places and at other times. Nature shows *reproducibility*. Nature shows no surprises. In fact, our memory and our thinking
Challenge 351 n are only possible because of this basic property of nature. (Can you confirm this?) As we will see, reproducibility leads to additional strong restrictions on the description of nature.

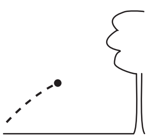
Without viewpoint-independence and reproducibility, talking to others or to oneself would be impossible. Even more importantly, we will discover that viewpoint-independence and reproducibility do more than determine the possibility of talking to each other: they also fix the *content* of what we can say to each other. In other words, we will see that our description of nature follows logically, almost without choice, from the simple fact that we can talk about nature to our friends.

Viewpoints

> Tolerance ... is the suspicion that the other might be right.
> Kurt Tucholski (1890–1935), German writer

> Tolerance – a strength one mainly wishes to political opponents.
> Wolfram Weidner (b. 1925) German journalist

When a young human starts to meet other people in childhood, it quickly finds out that certain experiences are shared, while others, such as dreams, are not. Learning to make



this distinction is one of the adventures of human life. In these pages, we concentrate on a section of the first type of experiences: *physical* observations. However, even among these, distinctions are to be made. In daily life we are used to assuming that weights, volumes, lengths and time intervals are independent of the viewpoint of the observer. We can talk about these observed quantities to anybody, and there are no disagreements over their values, provided they have been measured correctly. However, other quantities do depend on the observer. Imagine talking to a friend after he jumped from one of the trees along our path, while he is still falling downwards. He will say that the forest floor is approaching with high speed, whereas the observer below will maintain that the floor is stationary. Obviously, the difference between the statements is due to their different viewpoints. The velocity of an object (in this example that of the forest floor or of the friend himself) is thus a less symmetric property than weight or size. Not all observers agree on its value.

In the case of viewpoint-dependent observations, understanding is still possible with the help of a little effort: each observer can *imagine* observing from the point of view of the other, and *check* whether the imagined result agrees with the statement of the other.* If the statement thus imagined and the actual statement of the other observer agree, the observations are consistent, and the difference in statements is due only to the different viewpoints; otherwise, the difference is fundamental, and they cannot agree or talk. Using this approach, you can even argue whether human feelings, judgements, or tastes arise from fundamental differences or not.

Challenge 352 ny

The distinction between viewpoint-independent (invariant) and viewpoint-dependent quantities is an essential one. Invariant quantities, such as mass or shape, describe *intrinsic* properties, and quantities depending on the observer make up the *state* of the system. Therefore, we must answer the following questions in order to find a *complete* description of the state of a physical system:

- Which viewpoints are possible?
- How are descriptions transformed from one viewpoint to another?
- Which observables do these symmetries admit?
- What do these results tell us about motion?

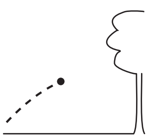
In the discussion so far, we have studied viewpoints differing in location, in orientation, in time and, most importantly, in motion. With respect to each other, observers can be at rest, move with constant speed, or accelerate. These 'concrete' changes of viewpoint are those we will study first. In this case the requirement of consistency of observations made by different observers is called the *principle of relativity*. The symmetries associated with this type of invariance are also called *external* symmetries. They are listed in Table 23.

Page 81
Page 184

A second class of fundamental changes of viewpoint concerns 'abstract' changes. Viewpoints can differ by the mathematical description used: such changes are called *changes of gauge*. They will be introduced first in the section on electrodynamics. Again, it is required that all statements be consistent across different mathematical descriptions. This requirement of consistency is called the *principle of gauge invariance*. The associated symmetries are called *internal* symmetries.

* Humans develop the ability to imagine that others can be in situations *different* from their own at the age of about four years. Therefore, before the age of four, humans are unable to conceive special relativity; afterwards, they can.

Ref. 141



The third class of changes, whose importance may not be evident from everyday life, is that of the behaviour of a system under exchange of its parts. The associated invariance is called *permutation symmetry*. It is a *discrete* symmetry, and we will encounter it in the second part of our adventure.

The three consistency requirements described above are called 'principles' because these basic statements are so strong that they almost completely determine the 'laws' of physics, as we will see shortly. Later on we will discover that looking for a complete description of the state of objects will also yield a complete description of their *intrinsic* properties. But enough of introduction: let us come to the heart of the topic.

Symmetries and groups

Since we are looking for a complete description of motion, we need to understand and describe the full set of symmetries of nature. A system is said to be symmetric or to possess a *symmetry* if it appears identical when observed from different viewpoints. We also say that the system possesses an *invariance* under change from one viewpoint to the other. Viewpoint changes are called *symmetry operations* or *transformations*. A symmetry is thus a transformation, or more generally, a set of transformations. However, it is more than that: the successive application of two symmetry operations is another symmetry operation. To be more precise, a symmetry is a set $G = \{a, b, c, ...\}$ of elements, the transformations, together with a binary operation $\circ$ called *concatenation* or *multiplication* and pronounced 'after' or 'times', in which the following properties hold for all elements a, b and c:

$$\begin{aligned} \textit{associativity}\text{, i.e. } \quad & (a \circ b) \circ c = a \circ (b \circ c) \\ \text{a } \textit{neutral element } e \text{ exists such that} \quad & e \circ a = a \circ e = a \\ \text{an } \textit{inverse element } a^{-1} \text{ exists such that} \quad & a^{-1} \circ a = a \circ a^{-1} = e \quad . \end{aligned} \tag{66}$$

Any set that fulfils these three defining properties, or axioms, is called a *(mathematical) group*. Historically, the notion of group was the first example of a mathematical structure which was defined in a completely abstract manner.* Can you give an example of a group
Challenge 353 n taken from daily life? Groups appear frequently in physics and mathematics, because
Ref. 142 symmetries are almost everywhere, as we will see.** Can you list the symmetry operations
Challenge 354 ny of the pattern of Figure 95?

* The term is due to Evariste Galois (1811–1832), the structure to Augustin-Louis Cauchy (1789–1857) and the axiomatic definition to Arthur Cayley (1821–1895).

** In principle, mathematical groups need not be symmetry groups; but it can be proven that all groups can be seen as transformation groups on some suitably defined mathematical space, so that in mathematics we can use the terms 'symmetry group' and 'group' interchangeably.

A group is called *Abelian* if its concatenation operation is commutative, i.e. if $a \circ b = b \circ a$ for all pairs of elements a and b. In this case the concatenation is sometimes called *addition*. Do rotations form an abelian group?

A subset $G_1 \subset G$ of a group G can itself be a group; one then calls it a *subgroup* and often says sloppily that G is *larger* than G_1 or that G is a *higher* symmetry group than G_1.

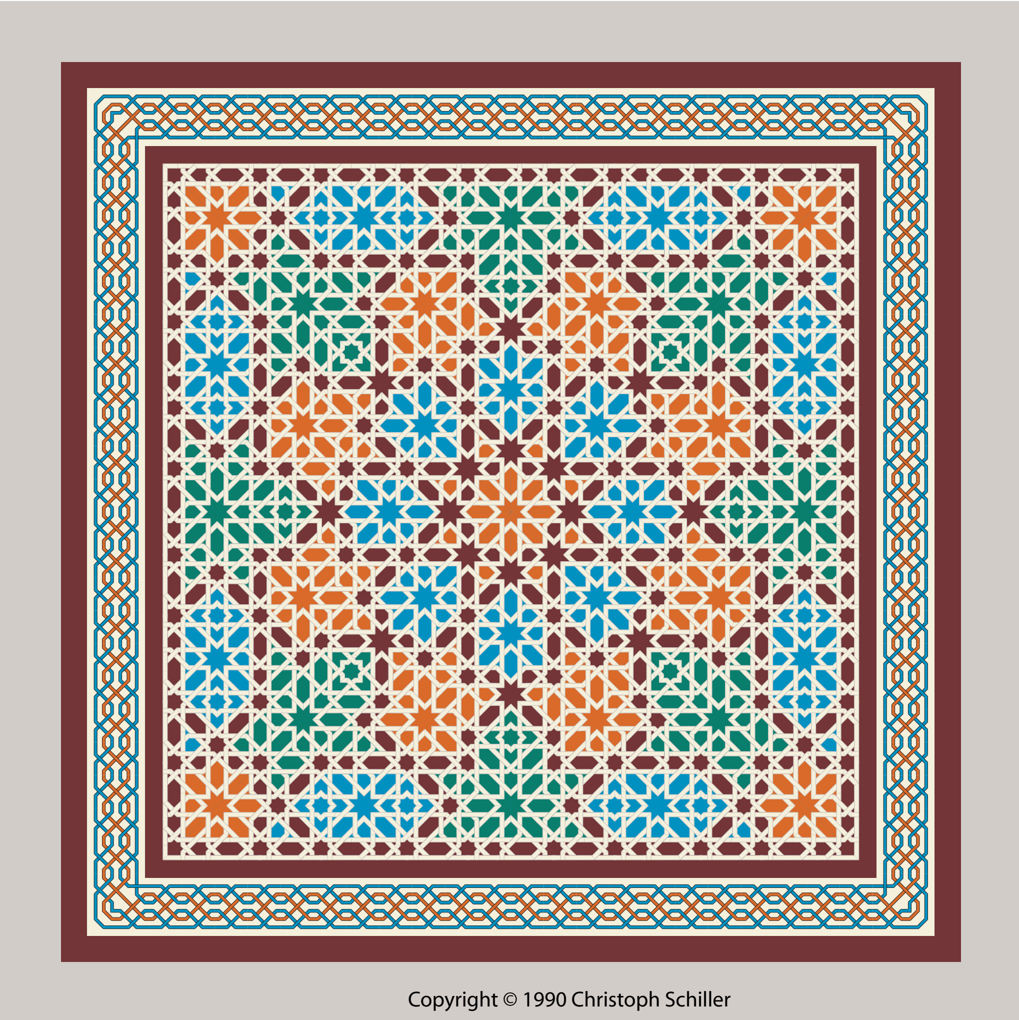


Figure 95 A Hispano–Arabic ornament from the Governor's Palace in Sevilla

Representations

Looking at a symmetric and composed system such as the one shown in Figure 95, we notice that each of its parts, for example each red patch, belongs to a set of similar objects, usually called a *multiplet*. Taken as a whole, the multiplet has (at least) the symmetry properties of the whole system. For some of the coloured patches in Figure 95 we need four objects to make up a full multiplet, whereas for others we need two, or only one, as in the case of the central star. In fact, in any symmetric system each part can be classified according to what type of multiplet it belongs to. Throughout our mountain ascent we will perform the same classification with every part of nature, with ever-increasing precision.

Challenge 355 e

A *multiplet* is a set of parts that transform into each other under all symmetry transformations. Mathematicians often call abstract multiplets *representations*. By specifying to which multiplet a component belongs, we describe in which way the component is part of the whole system. Let us see how this classification is achieved.

In mathematical language, symmetry transformations are often described by matrices. For example, in the plane, a reflection along the first diagonal is represented by the matrix

$$D(\text{refl}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} , \qquad (67)$$

since every point (x, y) becomes transformed to (y, x) when multiplied by the matrix $D(\text{refl})$. Therefore, for a mathematician a *representation* of a symmetry group G is an assignment of a matrix $D(a)$ to each group element a such that the representation of the concatenation of two elements a and b is the product of the representations D of the elements:

Challenge 356 e

$$D(a \circ b) = D(a)D(b) . \qquad (68)$$

For example, the matrix of equation (67), together with the corresponding matrices for all the other symmetry operations, have this property.*

For every symmetry group, the construction and classification of all possible representations is an important task. It corresponds to the classification of all possible multiplets a symmetric system can be made of. In this way, understanding the classification of all multiplets and parts which can appear in Figure 95 will teach us how to classify all possible parts of which an object or an example of motion can be composed!

A representation D is called *unitary* if all matrices $D(a)$ are unitary.** Almost all representations appearing in physics, with only a handful of exceptions, are unitary: this term is the most restrictive, since it specifies that the corresponding transformations are one-to-one and invertible, which means that one observer never sees more or less than

* There are some obvious, but important, side conditions for a representation: the matrices $D(a)$ must be invertible, or non-singular, and the identity operation of G must be mapped to the unit matrix. In even more compact language one says that a representation is a *homomorphism* from G into the group of non-singular or invertible matrices. A matrix D is invertible if its determinant $\det D$ is not zero.

In general, if a mapping f from a group G to another G' satisfies

$$f(a \circ_G b) = f(a) \circ_{G'} f(b) , \qquad (69)$$

the mapping f is called an *homomorphism*. A homomorphism f that is one-to-one (injective) and onto (surjective) is called a *isomorphism*. If a representation is also injective, it is called *faithful*, *true* or *proper*.

In the same way as groups, more complex mathematical structures such as rings, fields and associative algebras may also be represented by suitable classes of matrices. A representation of the field of complex numbers is given in Appendix D.

** The *transpose* A^T of a matrix A is defined element-by-element by $(A^T)_{ik} = A_{ki}$. The *complex conjugate* A^* of a matrix A is defined by $(A^*)_{ik} = (A_{ik})^*$. The *adjoint* $A^\dagger$ of a matrix A is defined by $A^\dagger = (A^T)^*$. A matrix is called *symmetric* if $A^T = A$, *orthogonal* if $A^T = A^{-1}$, *Hermitean* or *self-adjoint* (the two are synonymous in all physical applications) if $A^\dagger = A$ (Hermitean matrices have real eigenvalues), and *unitary* if $A^\dagger = A^{-1}$. Unitary matrices have eigenvalues of norm one. Multiplication by a unitary matrix is a one-to-one mapping; since the time evolution of physical systems is a mapping from one time to another, evolution is always described by a unitary matrix. A *real* matrix obeys $A^* = A$, an *antisymmetric* or *skew-symmetric* matrix is defined by $A^T = -A$, an *anti-Hermitean* matrix by $A^\dagger = -A$ and an *anti-unitary* matrix by $A^\dagger = -A^{-1}$. All the mappings described by these special types of matrices are one-to-one. A matrix is *singular*, i.e. not one-to-one, if $\det A = 0$.

another. Obviously, if an observer can talk to a second one, the second one can also talk to the first.

The final important property of a multiplet, or representation, concerns its structure. If a multiplet can be seen as composed of sub-multiplets, it is called *reducible*, else *irreducible*; the same is said about representations. The irreducible representations obviously cannot be decomposed any further. For example, the symmetry group of Figure 95, commonly called D_4, has eight elements. It has the general, faithful, unitary and irreducible matrix representation

Challenge 357 e

$$\begin{pmatrix} \cos n\pi/2 & -\sin n\pi/2 \\ \sin n\pi/2 & \cos n\pi/2 \end{pmatrix} \; n = 0..3, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}. \tag{70}$$

The representation is an *octet*. The complete list of possible irreducible representations of the group D_4 is given by *singlets*, *doublets* and *quartets*. Can you find them all? These representations allow the classification of all the white and black ribbons that appear in the figure, as well as all the coloured patches. The most symmetric elements are singlets, the least symmetric ones are members of the quartets. The complete system is always a singlet as well.

Challenge 358 ny

With these concepts we are ready to talk about motion with improved precision.

Symmetries, motion and Galilean physics

Every day we experience that we are able to talk to each other about motion. It must therefore be possible to find an *invariant* quantity describing it. We already know it: it is the *action*. Lighting a match is a change. It is the same whether it is lit here or there, in one direction or another, today or tomorrow. Indeed, the (Galilean) action is a number whose value is the same for each observer *at rest*, independent of his orientation or the time at which he makes his observation.

In the case of the Arabic pattern of Figure 95, the symmetry allows us to deduce the list of multiplets, or representations, that can be its building blocks. This approach must be possible for motion as well. We deduced the classification of the ribbons in the Arabic pattern into singlets, doublets, etc. from the various possible observation viewpoints. For a moving system, the building blocks, corresponding to the ribbons, are the *observables*. Since we observe that nature is symmetric under many different changes of viewpoint, we can classify all observables. To do so, we need to take the list of all viewpoint transformations and deduce the list of all their representations.

Our everyday life shows that the world stays unchanged after changes in position, orientation and instant of observation. One also speaks of space translation invariance, rotation invariance and time translation invariance. These transformations are different from those of the Arabic pattern in two respects: they are *continuous* and they are *unbounded*. As a result, their representations will generally be continuously variable and without bounds: they will be *quantities* or *magnitudes*. In other words, observables will be constructed with *numbers*. In this way we have deduced why numbers are *necessary* for any description of motion.*

* Only scalars, in contrast to vectors and higher-order tensors, may also be quantities which only take a

Table 22 Correspondences between the symmetries of an ornament, a flower and nature as a whole

System	Hispano–Arabic pattern	Flower	Motion
Structure and components	set of ribbons and patches	set of petals, stem	motion path and observables
System symmetry	pattern symmetry	flower symmetry	symmetry of Lagrangian
Mathematical description of the symmetry group	D_4	C_5	in Galilean relativity: position, orientation, instant and velocity changes
Invariants	number of multiplet elements	petal number	number of coordinates, magnitude of scalars, vectors and tensors
Representations of the components	multiplet types of elements	multiplet types of components	tensors, including scalars and vectors
Most symmetric representation	singlet	part with circular symmetry	scalar
Simplest faithful representation	quartet	quintet	vector
Least symmetric representation	quartet	quintet	no limit (tensor of infinite rank)

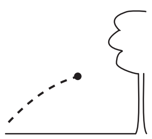
Since observers can differ in orientation, most representations will be objects possessing a direction. To cut a long story short, the symmetry under change of observation position, orientation or instant leads to the result that all observables are either 'scalars', 'vectors' or higher-order 'tensors.'*

A *scalar* is an observable quantity which stays the same for all observers: it corresponds to a singlet. Examples are the mass or the charge of an object, the distance between two points, the distance of the horizon, and many others. Their possible values are (usually) continuous, unbounded and without direction. Other examples of scalars are the potential at a point and the temperature at a point. Velocity is obviously not a scalar; nor is the coordinate of a point. Can you find more examples and counterexamples? Challenge 360 n

Energy is a puzzling observable. It is a scalar if only changes of place, orientation and instant of observation are considered. But energy is not a scalar if changes of observer speed are included. Nobody ever searched for a generalization of energy that is a scalar also for moving observers. Only Albert Einstein discovered it, completely by accident. More about this issue shortly.

Challenge 359 e discrete set of values, such as +1 or −1 only. In short, only scalars may be *discrete* observables.

* Later on, *spinors* will be added to, and complete, this list.



Any quantity which has a magnitude and a direction and which 'stays the same' with respect to the environment when changing viewpoint is a *vector*. For example, the arrow between two fixed points on the floor is a vector. Its length is the same for all observers; its direction changes from observer to observer, but not with respect to its environment. On the other hand, the arrow between a tree and the place where a rainbow touches the Earth is *not* a vector, since that place does not stay fixed with respect to the environment, when the observer changes.

Mathematicians say that vectors are directed entities staying invariant under coordinate transformations. Velocities of objects, accelerations and field strength are examples of Challenge 361 e vectors. (Can you confirm this?) The magnitude of a vector is a scalar: it is the same for any observer. By the way, a famous and baffling result of nineteenth-century experiments is that the velocity of light is *not* a vector for Galilean transformations. This mystery will be solved shortly.

Tensors are generalized vectors. As an example, take the moment of inertia of an object. Page 83 It specifies the dependence of the angular momentum on the angular velocity. For any object, doubling the magnitude of angular velocity doubles the magnitude of angular momentum; however, the two vectors are not parallel to each other if the object is not a Page 104 sphere. In general, if any two vector quantities are proportional, in the sense that doubling the magnitude of one vector doubles the magnitude of the other, but without the two vectors being parallel to each other, then the proportionality 'factor' is a (second order) *tensor*. Like all proportionality factors, tensors have a magnitude. In addition, tensors have a direction and a *shape*: they describe the connection between the vectors they relate. Just as vectors are the simplest quantities with a magnitude and a direction, so tensors are the simplest quantities with a magnitude and with a direction depending on a second, chosen direction. Vectors can be visualized as oriented arrows; tensors can be visualized Challenge 363 n as oriented ellipsoids.* Can you name another example of tensor?

Let us get back to the description of motion. Table 22 shows that in physical systems we always have to distinguish between the symmetry of the whole Lagrangian – corresponding to the symmetry of the complete pattern – and the representation of the observables – corresponding to the ribbon multiplets. Since the action must be a scalar, and since all observables must be tensors, Lagrangians contain sums and products of tensors only in combinations forming scalars. Lagrangians thus contain only scalar products or generalizations thereof. In short, Lagrangians always look like

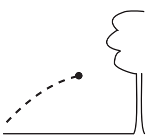
$$L = \alpha\, a_{\mathrm{i}} b^{\mathrm{i}} + \beta\, c_{\mathrm{jk}} d^{\mathrm{jk}} + \gamma\, e_{\mathrm{lmn}} f^{\mathrm{lmn}} + \dots \qquad (71)$$

where the indices attached to the variables a, b, c etc. always come in matching pairs to be

* A rank-n tensor is the proportionality factor between a rank-1 tensor, i.e. between a vector, and an rank-$(n-1)$ tensor. Vectors and scalars are rank 1 and rank 0 tensors. Scalars can be pictured as spheres, vectors as arrows, and rank-2 tensors as ellipsoids. Tensors of higher rank correspond to more and more complex shapes.

A vector has the same length and direction for every observer; a tensor (of rank 2) has the same determinant, the same trace, and the same sum of diagonal subdeterminants for all observers.

A vector is described mathematically by a *list* of components; a tensor (of rank 2) is described by a *matrix* of components. The rank or order of a tensor thus gives the number of indices the observable has. Can you Challenge 362 e show this?



summed over. (Therefore summation signs are usually simply left out.) The Greek letters represent constants. For example, the action of a free point particle in Galilean physics was given as

$$S = \int L \,\mathrm{d}t = \frac{m}{2} \int v^2 \,\mathrm{d}t \qquad (72)$$

which is indeed of the form just mentioned. We will encounter many other cases during our study of motion.*

Galileo already understood that motion is also invariant under change of viewpoints Page 81 with different velocity. However, the action just given does not reflect this. It took some years to find out the correct generalization: it is given by the theory of special relativity. But before we study it, we need to finish the present topic.

Reproducibility, conservation and Noether's theorem

> I will leave my mass, charge and momentum to science.
>
> Graffito

The reproducibility of observations, i.e. the symmetry under change of instant of time or 'time translation invariance', is a case of viewpoint-independence. (That is not obvi- Challenge 365 ny ous; can you find its irreducible representations?) The connection has several important consequences. We have seen that symmetry implies invariance. It turns out that for *continuous* symmetries, such as time translation symmetry, this statement can be made more precise: for any continuous symmetry of the Lagrangian there is an associated conserved constant of motion and vice versa. The exact formulation of this connection is the theorem of Emmy Noether.** She found the result in 1915 when helping Albert Einstein and

* By the way, is the usual list of possible observation viewpoints – namely different positions, different observation instants, different orientations, and different velocities – also *complete* for the action (72)? Sur- Ref. 143 prisingly, the answer is no. One of the first who noted this fact was Niederer, in 1972. Studying the quantum theory of point particles, he found that even the action of a Galilean free point particle is invariant under some additional transformations. If the two observers use the coordinates $(t, \mathbf{x})$ and (τ, ξ), the action (72) Challenge 364 ny is invariant under the transformations

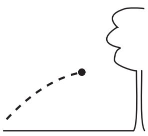
$$\xi = \frac{\mathbf{R}\mathbf{x} + \mathbf{x}_0 + \mathbf{v}t}{\gamma t + \delta} \quad \text{and} \quad \tau = \frac{\alpha t + \beta}{\gamma t + \delta} \quad \text{with} \quad \mathbf{R}^T\mathbf{R} = \mathbf{1} \quad \text{and} \quad \alpha\delta - \beta\gamma = 1\,. \qquad (73)$$

where $\mathbf{R}$ describes the rotation from the orientation of one observer to the other, $\mathbf{v}$ the velocity between the two observers, and $\mathbf{x}_0$ the vector between the two origins at time zero. This group contains two important special cases of transformations:

$$\text{The connected, static Galilei group } \xi = \mathbf{R}\mathbf{x} + \mathbf{x}_0 + \mathbf{v}t \quad \text{and} \quad \tau = t$$
$$\text{The transformation group SL(2,R) } \xi = \frac{\mathbf{x}}{\gamma t + \delta} \quad \text{and} \quad \tau = \frac{\alpha t + \beta}{\gamma t + \delta} \qquad (74)$$

The latter, three-parameter group includes *spatial inversion*, *dilations*, *time translation* and a set of time-dependent transformations such as $\xi = \mathbf{x}/t$, $\tau = 1/t$ called *expansions*. Dilations and expansions are rarely mentioned, as they are symmetries of point particles only, and do not apply to everyday objects and systems. They will return to be of importance later on, however.

** Emmy Noether (b. 1882 Erlangen, d. 1935 Bryn Mayr), German mathematician. The theorem is only



David Hilbert, who were both struggling and competing at constructing general relativity. Ref. 144 However, the result applies to any type of Lagrangian.

Noether investigated continuous symmetries depending on a continuous parameter b. A viewpoint transformation is a symmetry if the action S does not depend on the value of b. For example, changing position as

$$x \mapsto x + b \tag{75}$$

leaves the action

$$S_0 = \int T(v) - U(x)\, \mathrm{d}t \tag{76}$$

invariant, since $S(b) = S_0$. This situation implies that

$$\frac{\partial T}{\partial v} = p = \text{const} \quad ; \tag{77}$$

in short, symmetry under change of position implies conservation of momentum. The converse is also true.

Challenge 366 ny In the case of symmetry under shift of observation instant, we find

$$T + U = \text{const} \quad ; \tag{78}$$

in other words, time translation invariance implies constant energy. Again, the converse is also correct. One also says that energy and momentum are the *generators* of time and space translations.

The conserved quantity for a continuous symmetry is sometimes called the *Noether charge*, because the term *charge* is used in theoretical physics to designate conserved extensive observables. So, energy and momentum are Noether charges. 'Electric charge', 'gravitational charge' (i.e. mass) and 'topological charge' are other common examples. Challenge 367 n What is the conserved charge for rotation invariance?

We note that the expression 'energy is conserved' has several meanings. First of all, it means that the energy of a *single* free particle is constant in time. Secondly, it means that the total energy of any number of independent particles is constant. Finally, it means that the energy of a *system* of particles, i.e. including their interactions, is constant in time. Collisions are examples of the latter case. Noether's theorem makes all of these points at Challenge 368 e the same time, as you can verify using the corresponding Lagrangians.

But Noether's theorem also makes, or rather repeats, an even stronger statement: if energy were not conserved, time could not be defined. The whole description of nature requires the existence of conserved quantities, as we noticed when we introduced the Page 38 concepts of object, state and environment. For example, we defined objects as *permanent* entities, that is, as entities characterized by conserved quantities. We also saw that the Page 141 introduction of time is possible only because in nature there are 'no surprises'. Noether's

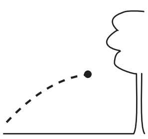
a sideline in her career which she dedicated mostly to number theory. The theorem also applies to gauge symmetries, where it states that to every gauge symmetry corresponds an identity of the equation of motion, and vice versa.

theorem describes exactly what such a 'surprise' would have to be: the non-conservation of energy. However, energy jumps have never been observed – not even at the quantum level.

Since symmetries are so important for the description of nature, Table 23 gives an overview of all the symmetries of nature we will encounter. Their main properties are also listed. Except for those marked as 'approximate' or 'speculative', an experimental proof of incorrectness of any of them would be a big surprise indeed.

Table 23 The symmetries of relativity and quantum theory with their properties; also the complete list of logical *inductions* used in the two fields

Symmetry	Type [number of parameters]	Space of action	Group topology	Possible representations	Conserved quantity	Vacuum/matter is symmetric	Main effect
Geometric or space-time, external, symmetries							
Time and space translation	$R \times R^3$ [4 par.]	space, time	not compact	scalars, vectors,	momentum and energy	yes/yes	allow everyday
Rotation	SO(3) [3 par.]	space	S^2	tensors	angular momentum	yes/yes	communication
Galilei boost	R^3 [3 par.]	space, time	not compact	scalars, vectors, tensors	velocity of centre of mass	yes/for low speeds	relativity of motion
Lorentz	homogeneous Lie SO(3,1) [6 par.]	space-time	not compact	tensors, spinors	energy-momentum $T^{\mu\nu}$	yes/yes	constant light speed
Poincaré ISL(2,C)	inhomogeneous Lie [10 par.]	space-time	not compact	tensors, spinors	energy-momentum $T^{\mu\nu}$	yes/yes	
Dilation invariance	R^+ [1 par.]	space-time	ray	n-dimen. continuum	none	yes/no	massless particles
Special conformal invariance	R^4 [4 par.]	space-time	R^4	n-dimen. continuum	none	yes/no	massless particles
Conformal invariance	[15 par.]	space-time	involved	massless tensors, spinors	none	yes/no	light cone invariance
Dynamic, interaction-dependent symmetries: gravity							



Symmetry	Type [number of parameters]	Space of action	Group topology	Possible representations	Conserved quantity	Vacuum/matter is symmetric	Main effect
$1/r^2$ gravity	SO(4) [6 par.]	config. space	as SO(4)	vector pair	perihelion direction	yes/yes	closed orbits
Diffeomorphism invariance	[∞ par.]	space-time	involved	space-times	local energy–momentum	yes/no	perihelion shift
Dynamic, classical and quantum-mechanical motion symmetries							
Motion(‘time’) inversion T	discrete	Hilbert or phase space	discrete	even, odd	T-parity	yes/no	reversibility
Parity(‘spatial’) inversion P	discrete	Hilbert or phase space	discrete	even, odd	P-parity	yes/no	mirror world exists
Charge conjugation C	global, antilinear, anti-Hermitean	Hilbert or phase space	discrete	even, odd	C-parity	yes/no	anti-particles exist
CPT	discrete	Hilbert or phase space	discrete	even	CPT-parity	yes/yes	makes field theory possible
Dynamic, interaction-dependent, gauge symmetries							
Electromagnetic classical gauge invariance	[∞ par.]	space of fields	un- important	un-important	electric charge	yes/yes	massless light
Electromagnetic q.m. gauge inv.	abelian Lie U(1) [1 par.]	Hilbert space	circle S^1	fields	electric charge	yes/yes	massless photon
Electromagnetic duality	abelian Lie U(1) [1 par.]	space of fields	circle S^1	abstract	abstract	yes/no	none
Weak gauge	non-abelian Lie SU(2) [3 par.]	Hilbert space	as $SU(3)$	particles	weak charge	no/ approx.	

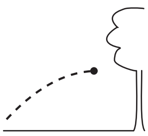
Symmetry	Type [number of parameters]	Space of action	Group topology	Possible representations	Conserved quantity	Vacuum/ matter is symmetric	Main effect
Colour gauge	non-abelian Lie SU(3) [8 par.]	Hilbert space	as $SU(3)$	coloured quarks	colour	yes/yes	massless gluons
Chiral symmetry	discrete	fermions	discrete	left, right	helicity	approximately	'massless' fermions[a]
Permutation symmetries							
Particle exchange	discrete	Fock space etc.	discrete	fermions and bosons	none	n.a./yes	Gibbs' paradox
Selected *speculative* symmetries of nature							
GUT	E_8, SO(10)	Hilbert	from Lie group	particles	from Lie group	yes/no	coupling constant convergence
N-supersymmetry[b]	global	Hilbert		particles, sparticles	T_{mn} and N spinors [c] Q_{imn}	no/no	'massless'[a] particles
R-parity	discrete	Hilbert	discrete	+1, -1	R-parity	yes/yes sfermions, gauginos	
Braid symmetry	discrete	own space	discrete	unclear	unclear	yes/maybe	unclear
Space-time duality	discrete	all	discrete	vacuum	unclear	yes/maybe	fixes particle masses
Event symmetry	discrete	space-time	discrete	nature	none	yes/no	unclear

For details about the connection between symmetry and induction, see page 625. The explanation of the terms in the table will be completed in the rest of the walk. The real numbers are denoted as R.

a. Only approximate; 'massless' means that $m \ll m_{Pl}$, i.e. that $m \ll 22\,\mu g$.

b. $N = 1$ supersymmetry, but not $N = 1$ supergravity, is probably a good approximation for nature at everyday energies.

c. $i = 1..N$.



In summary, since we can *talk* about nature we can deduce several of its symmetries, in particular its symmetry under time and space translations. From nature's symmetries, using Noether's theorem, we can deduce the conserved charges, such as energy or linear and angular momentum. In other words, the definition of mass, space and time, together with their symmetry properties, is *equivalent* to the conservation of energy and momentum. Conservation and symmetry are two ways to express the same property of nature. To put it simply, our ability to talk about nature means that energy and momentum are conserved.

In general, the most elegant way to uncover the 'laws' of nature is to search for nature's symmetries. In many historical cases, once this connection had been understood, physics made rapid progress. For example, Albert Einstein discovered the theory of relativity in this way, and Paul Dirac started off quantum electrodynamics. We will use the same method throughout our walk; in its third part we will uncover some symmetries which are even more mind-boggling than those of relativity. Now, though, we will move on to the next approach to a global description of motion.

Curiosities and fun challenges about motion symmetry

As diet for your brain, a few questions to ponder:

▪ What is the path followed by four turtles starting on the four angles of a square, if each of them continuously walks at the same speed towards the next one? Challenge 369 ny

▪ What is the symmetry of a simple oscillation? And of a wave? Challenge 370 n

▪ For what systems is motion reversal a symmetry transformation? Challenge 371 n

▪ What is the symmetry of a continuous rotation? Challenge 372 ny

▪ A sphere has a tensor for the moment of inertia that is diagonal with three equal numbers. The same is true for a cube. Can you distinguish spheres and cubes by their rotation behaviour? Challenge 373 ny

▪ Is there a motion in nature whose symmetry is perfect? Challenge 374 ny

Simple motions of extended bodies – oscillations and waves

We defined action, and thus change, as the integral of the Lagrangian, and the Lagrangian as the difference between kinetic and potential energy. One of the simplest systems in nature is a mass m attached to a spring. Its Lagrangian is given by

$$L = \frac{1}{2}mv^2 - kx^2 \ , \tag{79}$$

where k is a quantity characterizing the spring, the so-called spring constant. The Lagrangian is due to Robert Hooke, in the seventeenth century. Can you confirm it? Challenge 375 e

The motion that results from this Lagrangian is periodic, as shown in Figure 96. The Lagrangian describes the oscillation of the spring length. The motion is exactly the same as that of a long pendulum. It is called *harmonic motion*, because an object vibrating rapidly in this way produces a completely pure – or harmonic – musical sound. (The musical instrument producing the purest harmonic waves is the transverse flute. This instrument thus gives the best idea of how harmonic motion 'sounds'.) The graph of a

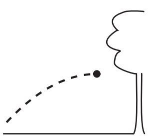


Table 24 Some mechanical frequency values found in nature

Observation	Frequency
Sound frequencies in gas emitted by black holes	*c.* 1 fHz
Precision in measured vibration frequencies of the Sun	down to 2 nHz
Vibration frequencies of the Sun	down to *c.* 300 nHz
Vibration frequencies that disturb gravitational radiation detection	down to 3 µHz
Lowest vibration frequency of the Earth Ref. 145	309 µHz
Resonance frequency of stomach and internal organs (giving the 'sound in the belly' experience)	1 to 10 Hz
Wing beat of tiny fly	*c.* 1000 Hz
Sound audible to young humans	20 Hz to 20 kHz
Sonar used by bats	up to over 100 kHz
Sonar used by dolphins	up to 150 kHz
Sound frequency used in ultrasound imaging	up to 15 MHz
Phonon (sound) frequencies measured in single crystals	up to 20 THz and more

harmonic or linear oscillation, shown in Figure 96, is called a *sine curve*; it can be seen as the basic building block of all oscillations. All other, non-harmonic oscillations in nature can be composed from sine curves, as we shall see shortly.

Page 189

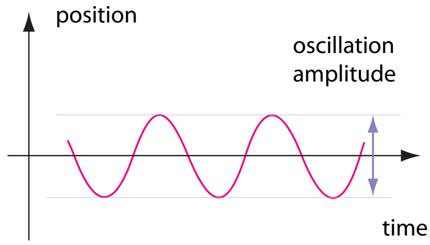


Figure 96 The simplest oscillation

Every oscillating motion continuously transforms kinetic energy into potential energy and vice versa. This is the case for the tides, the pendulum, or any radio receiver. But many oscillations also diminish in time: they are damped. Systems with large damping, such as the shock absorbers in cars, are used to avoid oscillations. Systems with *small* damping are useful for making precise and long-running clocks. The simplest measure of damping is the number of oscillations a system takes to reduce its amplitude to $1/e \approx 1/2.718$ times the original value. This characteristic number is the so-called *Q-factor*, named after the abbreviation of 'quality factor'. A poor Q-factor is 1 or less, an extremely good one is 100 000 or more. (Can you write down a simple Lagrangian for a damped oscillation with a given Q-factor?) In nature, damped oscillations do not usually keep constant frequency; however, for the simple pendulum this remains the case to a high degree of accuracy. The reason is that for a pendulum, the frequency does not depend significantly on the amplitude (as long as the amplitude is smaller than about 20°). This is one reason why pendulums are used as oscillators in mechanical clocks.

Challenge 376 ny

Obviously, for a good clock, the driving oscillation must not only show small damping, but must also be independent of temperature and be insensitive to other external influences. An important development of the twentieth century was the introduction of quartz crystals as oscillators. Technical quartzes are crystals of the size of a few grains of

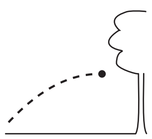


Figure yet to be included

Figure 97 Decomposing a general wave or signal into harmonic waves

sand; they can be made to oscillate by applying an electric signal. They have little temperature dependence and a large Q-factor, and therefore low energy consumption, so that precise clocks can now run on small batteries.

All systems that oscillate also emit waves. In fact, oscillations only appear in extended systems, and oscillations are only the simplest of motions of extended systems. The general motion of an extended system is the wave.

Waves and their motion

Waves are travelling imbalances, or, equivalently, travelling oscillations. Waves move, even though the substrate does not move. Every wave can be seen as a superposition of *harmonic* waves. Can you describe the difference in wave shape between a pure harmonic tone, a musical sound, a noise and an explosion? Every sound effect can be thought of as being composed of harmonic waves. Harmonic waves, also called *sine waves* or *linear waves*, are the building blocks of which all internal motions of an extended body are constructed.

Challenge 377 e

Every harmonic wave is characterized by an oscillation frequency and a propagation velocity. Low-amplitude water waves show this most clearly.

Waves appear inside all *extended* bodies, be they solids, liquids, gases or plasmas. Inside fluid bodies, waves are *longitudinal*, meaning that the wave motion is in the same direction as the wave oscillation. Sound in air is an example of a longitudinal wave. Inside solid bodies, waves can also be *transverse*; in that case the wave oscillation is perpendicular to the travelling direction.

Waves appear also on *interfaces* between bodies: water–air interfaces are a well-known case. Even a saltwater–freshwater interface, so-called *dead water*, shows waves: they can appear even if the upper surface of the water is immobile. Any flight in an aeroplane provides an opportunity to study the regular cloud arrangements on the interface between warm and cold air layers in the atmosphere. Seismic waves travelling along the boundary between the sea floor and the sea water are also well-known. General surface waves are usually neither longitudinal nor transverse, but of a mixed type.

On water surfaces, one classifies waves according to the force that restores the plane surface. The first type, *surface tension waves*, plays a role on scales up to a few centimetres. At longer scales, gravity takes over as the main restoring force and one speaks of *gravity waves*. This is the type we focus on here. Gravity waves in water, in contrast to surface

Table 25 Some wave velocities

Wave	Velocity
Tsunami	around 200 m/s
Sound in most gases	0.3 km/s
Sound in air at 273 K	331 m/s
Sound in air at 293 K	343 m/s
Sound in helium at 293 K	1.1 km/s
Sound in most liquids	1.1 km/s
Sound in water at 273 K	1.402 km/s
Sound in water at 293 K	1.482 km/s
Sound in gold	4.5 km/s
Sound in steel	5.790 km/s
Sound in granite	5.8 km/s
Sound in glass	5.9 km/s
Sound in beryllium	12.8 km/s
Sound in boron	up to 15 km/s
Sound in diamond	up to 18 km/s
Sound in fullerene (C_{60})	up to 26 km/s
Plasma wave velocity in InGaAs	600 km/s
Light in vacuum	$2.998 \cdot 10^8$ m/s

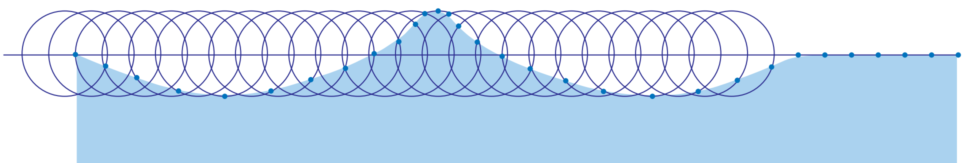


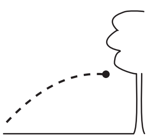
Figure 98 The formation of gravity waves on water

tension waves, are not sinusoidal. This is because of the special way the water moves in such a wave. As shown in Figure 98, the surface water moves in circles; this leads to the typical, asymmetrical wave shape with short sharp crests and long shallow troughs. (As long as there is no wind and the floor below the water is horizontal, the waves are also symmetric under front-to-back reflection.)

For water gravity waves, as for many other waves, the speed depends on the wavelength. Indeed, the speed c of water waves depends on the wavelength λ and on the depth of the water d in the following way:

$$c = \sqrt{\frac{g\lambda}{2\pi} \tanh \frac{2\pi d}{\lambda}} \;, \tag{80}$$

where g is the acceleration due to gravity (and an amplitude much smaller than the wavelength is assumed). The formula shows two limiting regimes. First, short or deep



waves appear when the water depth is larger than half the wavelength; for *deep waves*, the phase velocity is $c \approx \sqrt{g\lambda/2\pi}$, thus wavelength dependent, and the group velocity is about half the phase velocity. Shorter deep waves are thus slower. Secondly, shallow or *long waves* appear when the depth is less than 5% of the wavelength; in this case, $c \approx \sqrt{gd}$, there is no dispersion, and the group velocity is about the same as the phase velocity. The most impressive shallow waves are tsunamis, the large waves triggered by submarine earthquakes. (The Japanese name is composed of *tsu*, meaning harbour, and *nami*, meaning wave.) Since tsunamis are shallow waves, they show little dispersion and thus travel over long distances; they can go round the Earth several times. Typical oscillation times are between 6 and 60 minutes, giving wavelengths between 70 and 700 km and speeds in the open sea of 200 to 250 m/s, similar to that of a jet plane. Their amplitude on the open sea is often of the order of 10 cm; however, the amplitude scales with depth d as $1/d^4$ and heights up to 40 m have been measured at the shore. This was the order of magnitude of the large and disastrous tsunami observed in the Indian Ocean on 26 December 2004.

Challenge 378 e

Waves can also exist in empty space. Both light and gravity waves are examples. The exploration of electromagnetism and relativity will tell us more about their properties.

Any study of motion must include the study of wave motion. We know from experience that waves can hit or even damage targets; thus every wave carries energy and momentum, even though (on average) no matter moves along the wave propagation direction. The *energy* E of a wave is the sum of its kinetic and potential energy. The kinetic energy (density) depends on the temporal change of the displacement u at a given spot: rapidly changing waves carry a larger kinetic energy. The potential energy (density) depends on the gradient of the displacement, i.e. on its spatial change: steep waves carry a larger potential energy than shallow ones. (Can you explain why the potential energy does not depend on the displacement itself?) For harmonic waves propagating along the direction z, each type of energy is proportional to the square of its respective displacement change:

Challenge 379 n

Ref. 146

$$E \sim \left(\frac{\partial u}{\partial t}\right)^2 + v^2 \left(\frac{\partial u}{\partial z}\right)^2 . \tag{81}$$

How is the energy density related to the frequency?

Challenge 380 ny

The *momentum* of a wave is directed along the direction of wave propagation. The momentum value depends on both the temporal and the spatial change of displacement u. For harmonic waves, the momentum (density) P is proportional to the product of these two quantities:

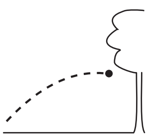
$$P_z \sim \frac{\partial u}{\partial t} \frac{\partial u}{\partial z} . \tag{82}$$

When two linear wave trains collide or interfere, the total momentum is conserved throughout the collision.

Waves, like moving bodies, carry energy and momentum. In simple terms, if you shout against a wall, the wall is hit. This hit, for example, can start avalanches on snowy mountain slopes. In the same way, waves, like bodies, can carry also angular momentum. (What type of wave is necessary for this to be possible?) However, we can distinguish six main properties that set the motion of waves apart from the motion of bodies.

Challenge 381 ny

- Waves can add up or cancel each other out; thus they can interpenetrate each other.



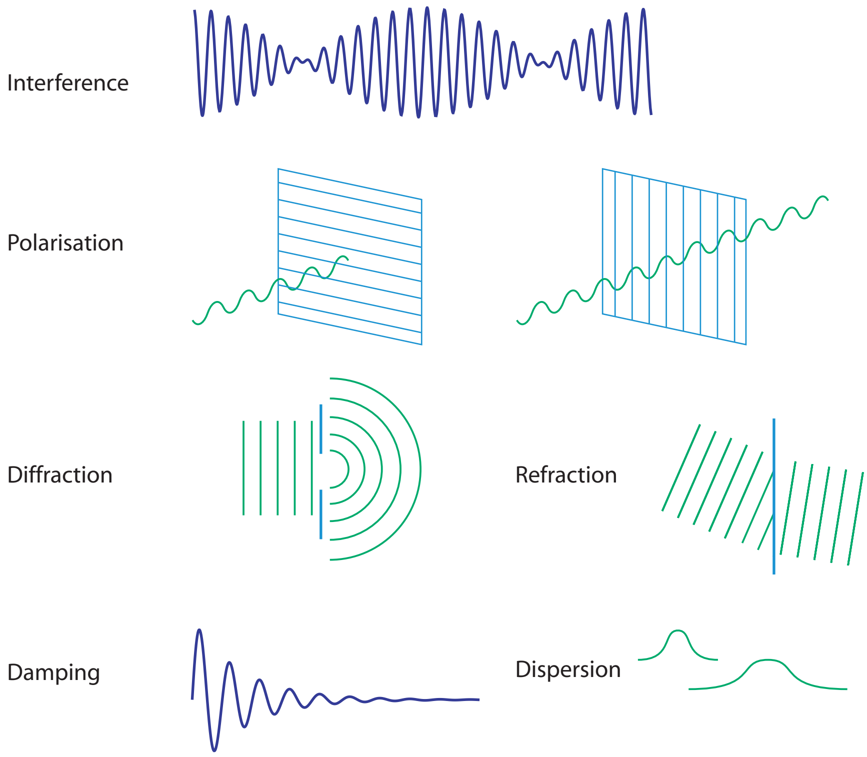


Figure 99 The six main properties of the motion of waves

These effects, called *superposition* and *interference*, are strongly tied to the linearity of most waves.

- Transverse waves in three dimensions can oscillate in different directions: they show *polarization*.
- Waves, such as sound, can go around corners. This is called *diffraction*.
- Waves change direction when they change medium. This is called *refraction*.
- Waves can have a frequency-dependent propagation speed. This is called *dispersion*.
- Often, the wave amplitude decreases over time: waves show *damping*.

Material bodies in everyday life do not behave in these ways when they move. These six wave effects appear because wave motion is the motion of *extended* entities. The famous debate whether electrons or light are waves or particles thus requires us to check whether these effects specific to waves can be observed or not. This is one topic of quantum theory. Before we study it, can you give an example of an observation that implies that a motion surely cannot be a wave?

Challenge 382 n

As a result of having a frequency f and a propagation velocity v, all sine waves are characterized by the distance λ between two neighbouring wave crests: this distance is called the wavelength. All waves obey the basic relation

$$\lambda f = v \ . \quad (83)$$

In many cases the wave velocity v depends on the wavelength of the wave. For example, this is the case for water waves. This change of speed with wavelength is called *dispersion*. In contrast, the speed of sound in air does not depend on the wavelength (to a high degree of accuracy). Sound in air shows almost no dispersion. Indeed, if there were dispersion for sound, we could not understand each other's speech at larger distances.

In everyday life we do not experience light as a wave, because the wavelength is only around one two-thousandth of a millimetre. But light shows all six effects typical of wave Page 516 motion. A rainbow, for example, can only be understood fully when the last five wave effects are taken into account. Diffraction and interference can even be observed with Challenge 383 n your fingers only. Can you tell how?

Like every anharmonic oscillation, every anharmonic wave can be decomposed into sine waves. Figure 97 gives examples. If the various sine waves contained in a disturbance propagate differently, the original wave will change in shape while it travels. That is the reason why an echo does not sound exactly like the original sound; for the same reason, a nearby thunder and a far-away one sound different.

All systems which oscillate also emit waves. Any radio or TV receiver contains oscillators. As a result, any such receiver is also a (weak) transmitter; indeed, in some countries the authorities search for people who listen to radio without permission listening to the radio waves emitted by these devices. Also, inside the human ear, numerous tiny structures, the hair cells, oscillate. As a result, the ear must also emit sound. This prediction, made in 1948 by Tommy Gold, was confirmed only in 1979 by David Kemp. These so-called *otoacoustic emissions* can be detected with sensitive microphones; they are presently being studied in order to unravel the still unknown workings of the ear and in order to Ref. 147 diagnose various ear illnesses without the need for surgery.

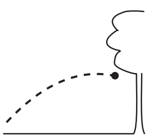
Since any travelling disturbance can be decomposed into sine waves, the term 'wave' is used by physicists for all travelling disturbances, whether they look like sine waves or not. In fact, the disturbances do not even have to be travelling. Take a standing wave: is it a wave or an oscillation? Standing waves do not travel; they are oscillations. But a standing wave can be seen as the superposition of two waves travelling in opposite directions. Since Challenge 384 ny all oscillations are standing waves (can you confirm this?), we can say that all oscillations are special forms of waves.

The most important travelling disturbances are those that are localized. Figure 97 shows an example of a localized wave group or pulse, together with its decomposition into harmonic waves. Wave groups are extensively used to talk and as signals for communication.

Why can we talk to each other? – Huygens' principle

The properties of our environment often disclose their full importance only when we ask simple questions. Why can we use the radio? Why can we talk on mobile phones? Why can we listen to each other? It turns out that a central part of the answer to these questions is that the space we live has an odd numbers of dimensions.

In spaces of even dimension, it is impossible to talk, because messages do not stop. This is an important result which is easily checked by throwing a stone into a lake: even after the stone has disappeared, waves are still emitted from the point at which it entered the water. Yet, when we stop talking, no waves are emitted any more.



– CS – text to be added – CS –

We can also say that Huygens' principle holds if the wave equation is solved by a circular wave leaving no amplitude behind it. Mathematicians translate this by requiring that the evolving delta function $\delta(c^2t^2 - r^2)$ satisfies the wave equation, i.e. that $\partial_t^2\delta = c^2\Delta\delta$. The delta function is that strange 'function' which is zero everywhere except at the origin, where it is infinite. A few more properties describe the precise way in which this happens.* It turns out that the delta function is a solution of the wave equation only if the space dimension is odd at least three.

In summary, the reason a room gets dark when we switch off the light, is that we live in a space with a number of dimensions which is odd and larger than one.

Signals

A signal is the transport of information. Every signal is motion of energy. Signals can be either objects or waves. A thrown stone can be a signal, as can a whistle. Waves are a more practical form of communication because they do not require transport of matter: it is easier to use electricity in a telephone wire to transport a statement than to send a messenger. Indeed, most modern technological advances can be traced to the separation between signal and matter transport. Instead of transporting an orchestra to transmit music, we can send radio signals. Instead of sending paper letters we write email messages. Instead of going to the library we browse the internet.

The greatest advances in communication have resulted from the use of signals to transport large amounts of energy. That is what electric cables do: they transport energy without transporting any matter. We do not need to attach our kitchen machines to the power station: we can get the energy via a copper wire.

For all these reasons, the term 'signal' is often meant to imply waves only. Voice, sound, electric signals, radio and light signals are the most common examples of wave signals.

Signals are characterized by their speed and their information content. Both quantities turn out to be limited. The limit on speed is the central topic of the theory of special
Page 249 relativity.

A simple limit on information content can be expressed when noting that the information flow is given by the detailed shape of the signal. The shape is characterized by a frequency (or wavelength) and a position in time (or space). For every signal – and every wave – there is a relation between the time-of-arrival error Δt and the frequency error Δf:

$$\Delta f \; \Delta t \geqslant \frac{1}{2} \; . \tag{84}$$

This indeterminacy relation expresses that, in a signal, it is impossible to specify both the time of arrival and the frequency with full precision. The two errors are (within a numerical factor) the inverse of each other. In fact, the indeterminacy relation is a feature of any wave phenomenon. You might want to test this relation with any wave in your
Challenge 385 e environment.

* The main property is $\int \delta x \mathrm{d}x = 1$. In mathematically precise terms, the delta 'function' is a distribution.

Figure to be included

Figure 100 The electrical signals measured in a nerve

Every indeterminacy relation is the consequence of a smallest entity. In the case of waves, the smallest entity of the phenomenon is the period (or cycle, as it used to be called). Whenever there is a smallest unit in a natural phenomenon, an indeterminacy relation results. Obviously, waves lead also to a *spatial* indeterminacy relation; can you write it down? (Challenge 386 ny) We will encounter other indeterminacy relations both in relativity and in quantum theory. As we will find out, they are due to smallest entities as well.

Whenever signals are sent, their content can be lost. Each of the six characteristics of waves listed on page 191 can lead to content degradation. Can you provide an example for each case? (Challenge 387 ny) The energy, the momentum and all other conserved properties of signals are never lost, of course. The disappearance of signals is akin to the disappearance of motion. When motion disappears by friction, it only seems to disappear, and is in fact transformed into heat. Similarly, when a signal disappears, it only seems to disappear, and is in fact transformed into noise. *(Physical) noise* is a collection of numerous disordered signals, in the same way that heat is a collection of numerous disordered movements.

All signal propagation is described by a wave equation. A famous example is the equation found by Hodgkin and Huxley ad a realistic approximation for the behaviour of electrical potential in nerves. (Ref. 148) Using facts about the behaviour of potassium and sodium ions, they found a elaborate equation that describes the voltage V in nerves, and thus the way the signals are propagated. The equation accurately describes the characteristic spikes measured in nerves, a shown in Figure 100. The figure clearly shows that these waves differ from sine waves: they are not harmonic. Anharmonicity is one result of nonlinearity. But nonlinearity can lead to even stronger effects.

Solitary waves and solitons

In August 1834, the Scottish engineer John Scott Russell (1808–1882) recorded a strange observation in a water canal in the countryside near Edinburgh. When a boat pulled through the channel was suddenly stopped, a strange water wave departed from it. It consisted of a single crest, about 10 m long and 0.5 m high, moving at about 4 m/s. He followed that crest with his horse for several kilometres: the wave died out only very slowly. He did not observe any dispersion, as is usual in water waves: the width of the crest remained constant. Russell then started producing such waves in his laboratory, and extensively studied their properties. (Ref. 153) He showed that the speed depended on the amplitude, in contrast to linear waves. The found that the depth d of the water canal was also an important parameter. In fact, the speed v, the amplitude A and the width L of these single-crested

Figure 101 A solitary water wave followed by a motor boat, reconstructing the discovery by Scott Russel (© Dugald Duncan)

Figure to be inserted

Figure 102 Solitons are stable against encounters

waves are related by

$$v = \sqrt{gd}\,(1 + \frac{A}{2d}) \quad \text{and} \quad L = \sqrt{\frac{4d^3}{3A}} \ . \tag{85}$$

As shown by these expressions, and noted by Russell, high waves are narrow and fast, whereas shallow waves are slow and wide. The shape of the waves is fixed during their motion. Today, these and all other stable waves with a single crest are called *solitary waves*. They appear only where the dispersion and the nonlinearity of the system exactly compensate for each other. Russell also noted that the solitary waves in water channels can cross each other unchanged, even when travelling in opposite directions; solitary waves with this property are called *solitons*. Solitons are stable against encounters, whereas solitary waves in general are not.

Only sixty years later, in 1895, Korteweg and de Vries found out that solitary waves in water channels have a shape described by

$$u(x,t) = A\,\mathrm{sech}^2\frac{x - vt}{L} \quad \text{where} \quad \mathrm{sech}\,x = \frac{2}{e^x + e^{-x}}\ , \tag{86}$$

and that the relation found by Russell was due to the wave equation

$$\frac{1}{\sqrt{gd}}\frac{\partial u}{\partial t} + \Big(1 + \frac{3}{2d}u\Big)\frac{\partial u}{\partial x} + \frac{d^2}{6}\frac{\partial^3 u}{\partial x^3} = 0\ . \tag{87}$$

This equation for the elongation u is called the *Korteweg–de Vries equation* in their honour.* The surprising stability of the solitary solutions is due to the opposite effect of the two terms that distinguish the equation from linear wave equations: for the solitary solutions, the nonlinear term precisely compensates for the dispersion induced by the third-derivative term.

For many decades such solitary waves were seen as mathematical and physical curiosities. But almost a hundred years later it became clear that the Korteweg–de Vries equation is a universal model for weakly nonlinear waves in the weak dispersion regime, and thus of basic importance. This conclusion was triggered by Kruskal and Zabusky, who in 1965 proved mathematically that the solutions (86) are unchanged in collisions. This discovery Ref. 151 prompted them to introduce the term *soliton*. These solutions do indeed interpenetrate one another without changing velocity or shape: a collision only produces a small positional shift for each pulse.

Solitary waves play a role in many examples of fluid flows. They are found in ocean currents; and even the red spot on Jupiter, which was a steady feature of Jupiter photographs for many centuries, is an example.

Solitary waves also appear when extremely high-intensity sound is generated in solids. Ref. 152 In these cases, they can lead to sound pulses of only a few nanometres in length. Solitary light pulses are also used inside certain optical communication fibres, where they provide Ref. 153 (almost) lossless signal transmission.

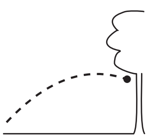
Towards the end of the twentieth century a second wave of interest in the mathematics of solitons arose, when quantum theorists became interested in them. The reason is simple but deep: a soliton is a 'middle thing' between a particle and a wave; it has features of both concepts. For this reason, solitons are now an essential part of any description of elementary particles, as we will find out later on.

Curiosities and fun challenges about waves and extended bodies

Oscillations, waves and signals are a limitless source of fascination.

▪ An orchestra is playing music in a large hall. At a distance of 30 m, somebody is listening to the music. At a distance of 3000 km, another person is listening to the music Challenge 388 ny via the radio. Who hears the music first?

* The equation can be simplified by transforming the variable u; most concisely, it can be rewritten as $u_t + u_{xxx} = 6uu_x$. As long as the solutions are sech functions, this and other transformed versions of the equation are known by the same name.



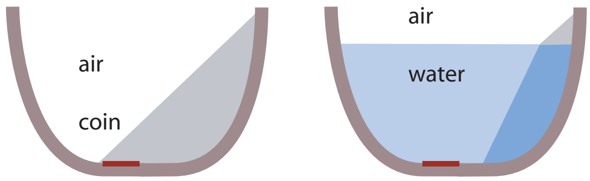


Figure 103 Shadows and refraction

▪ What is the period of a simple pendulum, i.e. a mass m attached to a massless string of length l? What is the period if the string is much longer than the radius of the Earth? Challenge 389 ny

▪ What path is followed by a body moving in a plane, but attached by a spring to a fixed point on the plane? Challenge 390 ny

▪ Light is a wave, as we will discover later on. As a result, light reaching the Earth from space is refracted when it enters the atmosphere. Can you confirm that as a result, stars appear somewhat higher in the night sky than they really are? Challenge 391 e

▪ What are the highest sea waves? This question has been researched systematically only recently, using satellites. The surprising result is that sea waves with a height of 25 m and more are *common*: there are a few such waves on the oceans at any given time. This result confirms the rare stories of experienced ship captains and explains many otherwise ship sinkings.

Surfers may thus get many chances to ride 30 m waves. (The record is just below this size.) But maybe the most impressive waves to surf are those of the Pororoca, a series of 4 m waves that move from the sea into the Amazonas River every spring, against the flow of the river. These waves can be surfed for tens of kilometres.

▪ All waves are damped, eventually. This effect is often frequency-dependent. Can you provide a confirmation of this dependence in the case of sound in air? Challenge 392 n

▪ When you make a hole with a needle in black paper, the hole can be used as a magnifying lens. (Try it.) Diffraction is responsible for the lens effect. By the way, the diffraction of light by holes was noted by Francesco Grimaldi in the seventeenth century; he deduced that light is a wave. His observations were later discussed by Newton, who wrongly dismissed them. Challenge 393 e

▪ Put a empty cup near a lamp, in such a way that the bottom of the cup remains in the shadow. When you fill the cup with water, some of the bottom will be lit, because of the refraction of the light from the lamp. The same effect allows us to build lenses. The same effect is at the basis of instruments such as the telescope. Page 525

▪ Are water waves transverse or longitudinal? Challenge 394 n

▪ The speed of water waves limits the speeds of ships. A surface ship cannot travel (much) faster than about $v_{\text{crit}} = \sqrt{0.16 g l}$, where $g = 9.8\,\text{m/s}^2$, l is its length, and 0.16 is a number determined experimentally, called the critical Froude number. This relation is valid for all vessels, from large tankers ($l = 100$ m gives $v_{\text{crit}} = 13$ m/s) down to ducks ($l = 0.3$ m gives $v_{\text{crit}} = 0.7$ m/s). The critical speed is that of a wave with the same wavelength as the ship. In fact, moving at higher speeds than the critical value is possible, but requires much more energy. (A higher speed is also possible if the ship surfs on a wave.) Therefore all water animals and ships are faster when they swim below the surface – where the limit due to surface waves does not exist – than when they swim on the surface. For example, ducks can swim three times as fast under water than on the surface.

Challenge 395 n

How far away is the *olympic swimming* record from the critical value?

▪ The group velocity of water waves (in deep water) is less than the velocity of the individual waves. As a result, when a group of wave crests travels, within the group the crests move from the back to the front, appearing at the back, travelling forward and then dying out at the front.

▪ One can hear the distant sea or a distant highway more clearly in the evening than in the morning. This is an effect of refraction. Sound speed decreases with temperature. In the evening, the ground cools more quickly than the air above. As a result, sound leaving the ground and travelling upwards is refracted downwards, leading to the long hearing distance. In the morning, usually the air is cold above and warm below. Sound is refracted upwards, and distant sound does not reach a listener on the ground. Refraction thus implies that mornings are quiet, and that one can hear more distant sounds in the evenings. Elephants use the sound situation during evenings to communicate over distances of more than 10 km. (They also use sound waves in the ground to communicate, but that is another story.)

▪ Refraction also implies that there is a sound channel in the ocean, and in the atmosphere. Sound speed decreases with temperature, and increases with pressure. At an ocean depth of 1 km, or at an atmospheric height of 13 to 17 km (that is at the top of the tallest cumulonimbus clouds or equivalently, at the middle of the ozone layer) sound has minimal speed. As a result, sound that starts from that level and tries to leave is channelled back to it. Whales use the sound channel to communicate with each other with

Challenge 396 e

beautiful songs; one can find recordings of these songs on the internet. The military successfully uses microphones placed at the sound channel in the ocean to locate submarines,

Ref. 149

and microphones on balloons in the atmospheric channel to listen for nuclear explosions. (In fact, sound experiments conducted by the military are the main reason why whales are deafened and lose their orientation, stranding on the shores. Similar experiments in the air with high-altitude balloons are often mistaken for flying saucers, as in the famous Roswell incident.)

Ref. 150

▪ Much smaller also animals communicate by sound waves. In 2003, it was found that herring communicate using noises they produce when farting. When they pass wind, the gas creates a ticking sound whose frequency spectrum reaches up to 20 kHz. One can even listen to recordings of this sound on the internet. The details of the communication, such as the differences between males and females, are still being investigated. It is possible that the sounds may also be used by predators to detect herring, and they might even by used by future fishing vessels.

▪ On windy seas, the white wave crests have several important effects. The noise stems from tiny exploding and imploding water bubbles. The noise of waves on the open sea is thus the superposition of many small explosions. At the same time, white crests are the events where the seas absorb carbon dioxide from the atmosphere, and thus reduce global warming.

Challenge 397 n

▪ Why are there many small holes in the ceilings of many office buildings?

▪ Which quantity determines the wavelength of water waves when a stone is thrown

Challenge 398 ny

into a pond?

Ref. 2

▪ Yakov Perelman lists the following four problems in his delightful physics problem book.

(1) A stone falling into a lake produces circular waves. What is the shape of waves produced by a stone falling into a river, where the water flows in one direction? Challenge 399 n

(2) It is possible to build a lens for sound, in the same way as it is possible to build lenses for light. What would such a lens look like? Challenge 400 n

(3) What is the sound heard inside a shell? Challenge 401 ny

(4) Light takes about eight minutes to travel from the Sun to the Earth. What consequence does this have for a sunrise? Challenge 402 n

▪ Can you describe how a Rubik's Cube is built? And its generalizations to higher numbers of segments? Is there a limit to the number of segments? Challenge 403 n These puzzles are even tougher than the search for a rearrangement of the cube. Similar puzzles can be found in the study of many mechanisms, from robots to textile machines.

▪ Typically, sound produces a pressure variation of 10^{-8} bar on the ear. How is this determined? Challenge 404 ny

The ear is indeed a sensitive device. It is now known that most cases of sea mammals, like whales, swimming onto the shore are due to ear problems: usually some military device (either sonar signals or explosions) has destroyed their ear so that they became deaf and lose orientation.

▪ *Infrasound*, inaudible sound below 20 Hz, is a modern topic of research. In nature, infrasound is emitted by earthquakes, volcanic eruptions, wind, thunder, waterfalls, falling meteorites and the surf. Glacier motion, seaquakes, avalanches and geomagnetic storms also emit infrasound. Ref. 154 Human sources include missile launches, traffic, fuel engines and air compressors.

It is known that high intensities of infrasound lead to vomiting or disturbances of the sense of equilibrium (140 dB or more for 2 minutes), and even to death (170 dB for 10 minutes). The effects of lower intensities on human health are not yet known.

Infrasound can travel several times around the world before dying down, as the explosion of the Krakatoa volcano showed in 1883. With modern infrasound detectors, sea surf can be detected hundreds of kilometres away. Infrasound detectors are even used to count meteorites at night. Very rarely, meteorites can be heard with the human ear.

▪ The method used to deduce the sine waves contained in a signal, as shown in Figure 97, is called the Fourier transformation. It is of importance throughout science and technology. In the 1980s, an interesting generalization became popular, called the *wavelet transformation*. In contrast to Fourier transformations, wavelet transformations allow us to localize signals in time. Wavelet transformations are used to compress digitally stored images in an efficient way, to diagnose aeroplane turbine problems, and in many other applications. Ref. 155

▪ If you like engineering challenges, here is one that is still open. How can one make a robust and efficient system that transforms the energy of sea waves into electricity? Challenge 405 r

▪ In our description of extended bodies, we assumed that each spot of a body can be followed separately throughout its motion. Is this assumption justified? What would happen if it were not? Challenge 406 r

▪ Bats fly at night using *echolocation*. Dolphins also use it. Sonar, used by fishing vessels to look for fish, copies the system of dolphins. Less well known is that humans have the same ability. Ref. 156 Have you ever tried to echolocate a wall in a completely dark room? You will be surprised at how easily this is possible. Just make a loud hissing or whistling noise that stops abruptly, and listen to the echo. You will be able to locate walls reliably. Challenge 407 e

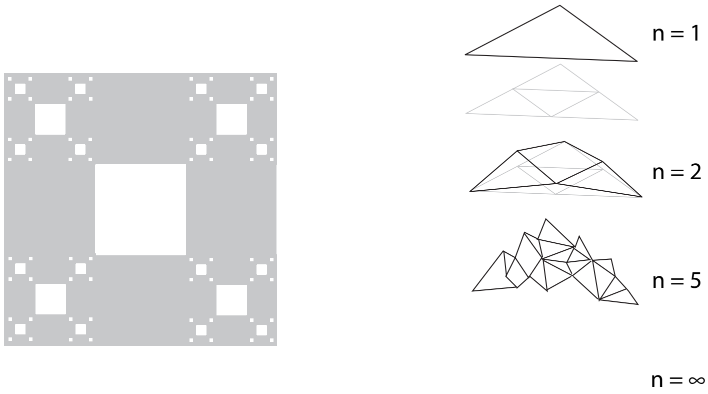


Figure 104 Floors and mountains as fractals

▪ A bicycle chain is an extended object with no stiffness. However, if it is made to rotate rapidly, it gets dynamical stiffness, and can roll down an inclined plane. This surprising effect can be seen on the http://www.iwf.de/Navigation/Projekte/LNW/Pohl/index.asp website.

Do extended bodies exist?

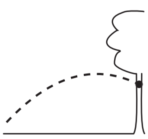
We have just discussed the motion of extended bodies in some detail. We have seem that extended bodies show wave motion. But are extended bodies found in nature? Strangely enough, this question has been one of the most intensely discussed questions in physics. Over the centuries, it has reappeared again and again, at each improvement of the description of motion; the answer has alternated between the affirmative and the negative. Many thinkers have been imprisoned, and many still are being persecuted, for giving answers that are not politically correct! In fact, the issue already arises in everyday life.

Mountains and fractals

Whenever we climb a mountain, we follow the outline of its shape. We usually describe this outline as a curved two-dimensional surface. In everyday life we find that this is a good approximation. But there are alternative possibilities. The most popular is the idea that mountains are fractal surfaces. A *fractal* was defined by Benoit Mandelbrot as a set that is self-similar under a countable but infinite number of magnification values.*
Page 52 We have already encountered fractal lines. An example of an algorithm for building a (random) fractal *surface* is shown on the right side of Figure 104. It produces shapes which
Ref. 157 look remarkably similar to real mountains. The results are so realistic that they are used in Hollywood movies. If this description were correct, mountains would be extended, but not continuous.

But mountains could also be fractals of a different sort, as shown in the left side of Figure 104. Mountain surfaces could have an infinity of small and smaller holes. In fact,

* For a definition of uncountability, see page 602.



one could also imagine that mountains are described as three-dimensional versions of the left side of the figure. Mountains would then be some sort of mathematical Swiss cheese. Can you devise an experiment to decide whether fractals provide the correct description for mountains? To settle the issue, a chocolate bar can help.

Challenge 408 n

Can a chocolate bar last forever?

> From a drop of water a logician could predict an Atlantic or a Niagara.
>
> Arthur Conan Doyle (1859–1930), *A Study in Scarlet*

Any child knows how to make a chocolate bar last forever: eat half the remainder every day. However, this method only works if matter is scale-invariant. In other words, the method only works if matter is either *fractal*, as it then would be scale-invariant for a discrete set of zoom factors, or *continuous*, in which case it would be scale-invariant for any zoom factor. Which case, if either, applies to nature?

Page 51

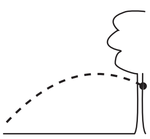
We have already encountered a fact making continuity a questionable assumption: continuity would allow us, as Banach and Tarski showed, to multiply food and any other matter by clever cutting and reassembling. Continuity would allow children to eat the *same* amount of chocolate every day, without ever buying a new bar. Matter is thus not continuous. Now, fractal chocolate is not ruled out in this way; but other experiments settle the question. Indeed, we note that melted materials do not take up much smaller volumes than solid ones. We also find that even under the highest pressures, materials do not shrink. Thus matter is not a fractal. What then is its structure?

To get an idea of the structure of matter we can take fluid chocolate, or even just some oil – which is the main ingredient of chocolate anyway – and spread it out over a large surface. For example, we can spread a drop of oil onto a pond on a day without rain or wind; it is not difficult to observe which parts of the water are covered by the oil and which are not. A small droplet of oil cannot cover a surface larger than – can you guess the value? Trying to spread the film further inevitably rips it apart. The child's method of prolonging chocolate thus does not work for ever: it comes to a sudden end. The oil experiment shows that there is a *minimum* thickness of oil films, with a value of about 2 nm. This simple experiment can even be conducted at home; it shows that there is a smallest size in matter. Matter is made of tiny components. This confirms the observations made by Joseph Loschmidt* in 1865, who was the first person to measure the size of the components of matter.** In 1865, it was not a surprise that matter was made of small

Challenge 409 n

* Joseph Loschmidt (b. 1821 Putschirn, d. 1895 Vienna) Austrian chemist and physicist. The oil experiment was popularized a few decades later, by Kelvin. It is often claimed that Benjamin Franklin was the first to conduct the oil experiment; that is wrong. Franklin did not measure the thickness, and did not even consider the question of the thickness. He did pour oil on water, but missed the most important conclusion that could be drawn from it. Even geniuses do not discover everything.

** Loschmidt knew that the (dynamic) viscosity of a gas was given by $\eta = \rho l v/3$, where ρ is the density of the gas, v the average speed of the components and l their mean free path. With Avogadro's prediction (made in 1811 without specifying any value) that a volume V of any gas always contains the same number N of components, one also has $l = V/\sqrt{2\pi N\sigma^2}$, where σ is the cross section of the components. (The cross section is the area of the shadow of an object.) Loschmidt then assumed that when the gas is liquefied, the



components, as the existence of a smallest size – but not its value – had already been deduced by Galileo, when studying some other simple questions.*

How high can animals jump?

Fleas can jump to heights a hundred times their size, humans only to heights about their own size. In fact, biological studies yield a simple observation: most animals, regardless of their size, achieve about the same jumping height of between 0.8 and 2.2 m, whether they are humans, cats, grasshoppers, apes, horses or leopards. The explanation of this fact takes only two lines. Can you find it?

Ref. 158

Challenge 410 n

The above observation seems to be an example of scale invariance. But there are some interesting exceptions at both ends of the mass range. At the small end, mites and other small insects do not achieve such heights because, like all small objects, they encounter the problem of air resistance. At the large end, elephants do not jump that high, because doing so would break their bones. But why do bones break at all?

Why are all humans of about the same size? Why are there no giant adults with a height of ten metres? Why aren't there any land animals larger than elephants? The answer yields the key to understanding the structure of matter. In fact, the materials of which we are made would not allow such changes of scale, as the bones of giants would collapse under the weight they have to sustain. Bones have a finite strength because their constituents stick to each other with a finite attraction. Continuous matter – which exists only in cartoons – could not break at all, and fractal matter would be infinitely fragile. Matter breaks under finite loads because it is composed of small basic constituents.

volume of the liquid is the sum of the volumes of the particles. He then measured all the involved quantities and determined N. The modern value of N, called *Avogadro's number* or *Loschmidt's number*, is $6.02 \cdot 10^{23}$ particles in 22.4 l of any gas at standard conditions (today called 1 mol).

* Galileo was brought to trial because of his ideas about atoms, not about the motion of the Earth, as is often claimed. To get a clear view of the matters of dispute in the case of Galileo, especially those of interest to physicists, the best text is the excellent book by Pietro Redondi, *Galileo eretico*, Einaudi, 1983, translated into English as *Galileo Heretic*, Princeton University Press, 1987. It is also available in many other languages. Redondi, a renowned historical scholar and colleague of Pierre Costabel, tells the story of the dispute between Galileo and the reactionary parts of the Catholic Church. He discovered a document of that time – the anonymous denunciation which started the trial – that allowed him to show that the condemnation of Galileo to life imprisonment for his views on the Earth's motion was organized by his friend the Pope to *protect* him from a sure condemnation to death over a different issue.

The reasons for his arrest, as shown by the denunciation, were not his ideas on astronomy and on the motion of the Earth, but his statements on matter. Galileo defended the view that since matter is not scale invariant, it must be made of 'atoms' or, as he called them, *piccolissimi quanti* – smallest quanta. This was and still is a heresy. A true Catholic is still not allowed to believe in atoms. Indeed, the theory of atoms is not compatible with the change of bread and wine into human flesh and blood, called *transsubstantiation*, which is a central tenet of the Catholic faith. In Galileo's days, church tribunals punished heresy, i.e. deviating personal opinions, by the death sentence. Despite being condemned to prison in his trial, Galileo published his last book, written as an old man under house arrest, on the scaling issue. Today, the Catholic Church still refuses to publish the proceedings and other documents of the trial. Its officials carefully avoid the subject of atoms, as any statement on this subject would make the Catholic Church into a laughing stock. In fact, quantum theory, named after the term used by Galileo, has become the most precise description of nature yet.

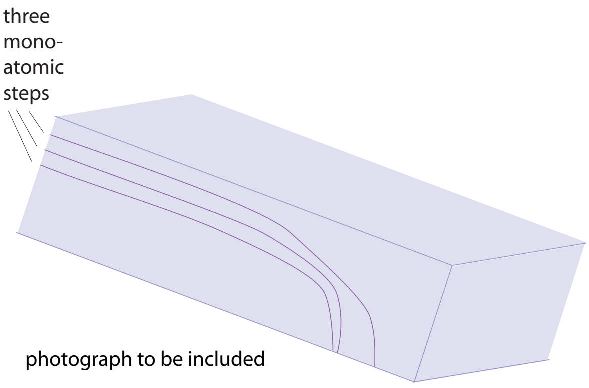


Figure 105 Atomic steps in broken gallium arsenide crystals can be seen under a light microscope

Felling trees

The gentle lower slopes of Motion Mountain are covered by trees. Trees are fascinating structures. Take their size. Why do trees have limited size? Already in the sixteenth century, Galileo knew that it is not possible to increase tree height without limits: at some point a tree would not have the strength to support its own weight. He estimated the maximum height to be around 90 m; the actual record, unknown to him at the time, seems to be 150 m, for the Australian tree *Eucalyptus regnans*. But why does a limit exist at all? The answer is the same as for bones: wood has a finite strength because it is not scale invariant; and it is not scale invariant because it is made of small constituents, namely atoms.*

In fact, the derivation of the precise value of the height limit is more involved. Trees must not break under strong winds. Wind resistance limits the height-to-thickness ratio

Challenge 411 ny h/d to about 50 for normal-sized trees (for $0.2\,\mathrm{m} < d < 2\,\mathrm{m}$). Can you say why? Thinner trees are limited in height to less than 10 m by the requirement that they return to the

Ref. 160 vertical after being bent by the wind.

Such studies of natural constraints also answer the question of why trees are made from wood and not, for example, from steel. You could check for yourself that the maximum height of a column of a given mass is determined by the ratio E/ρ^2 between the elastic

Challenge 412 n module and the square of the mass density. Wood is actually the material for which this

Ref. 161 ratio is highest. Only recently have material scientists managed to engineer slightly better ratios with fibre composites.

Why do materials break at all? All observations yield the same answer and confirm Galileo's reasoning: because there is a smallest size in materials. For example, bodies under stress are torn apart at the position at which their strength is minimal. If a body were completely homogeneous, it could not be torn apart; a crack could not start anywhere. If a body had a fractal Swiss-cheese structure, cracks would have places to start, but they would need only an infinitesimal shock to do so.

Ref. 159 * There is another important limiting factor: the water columns inside trees must not break. Both factors seem to yield similar limiting heights.

It is not difficult to confirm experimentally the existence of smallest size in solids. It is sufficient to break a single crystal, such as a gallium arsenide wafer, in two. The breaking surface is either completely flat or shows extremely small steps, as shown in Figure 105. These steps are visible under a normal light microscope. It turns out that all the step heights are multiples of a smallest height: its value is about 0.2 nm. The existence of a smallest height, corresponding to the height of an atom, contradicts all possibilities of scale invariance in matter.

The sound of silence

Climbing the slopes of Motion Mountain, we arrive in a region of the forest covered with deep snow. We stop for a minute and look around. It is dark; all the animals are asleep; there is no wind and there are no sources of sound. We stand still, without breathing, and listen to the silence. (You can have this experience also in a sound studio such as those used for musical recordings, or in a quiet bedroom at night.) In situations of complete silence, the ear automatically becomes more sensitive*; we then have a strange experience. We hear two noises, a lower- and a higher-pitched one, which are obviously generated inside the ear. Experiments show that the higher note is due to the activity of the nerve cells in the inner ear. The lower note is due to pulsating blood streaming through the head. But why do we hear a noise at all?

Many similar experiments confirm that whatever we do, we can never eliminate noise from measurements. This unavoidable type of noise is called *shot noise* in physics. The statistical properties of this type of noise actually correspond precisely to what would be expected if flows, instead of being motions of continuous matter, were transportation of a large number of equal, small and discrete entities. Thus, simply listening to noise proves that electric current is made of electrons, that air and liquids are made of molecules, and that light is made of photons. In a sense, the sound of silence is the sound of atoms. Shot noise would not exist in continuous systems.

Little hard balls

> I prefer knowing the cause of a single thing to being king of Persia.
>
> Democritus

Precise observations show that matter is neither continuous nor a fractal: matter is made of smallest basic particles. Galileo, who deduced their existence by thinking about giants and trees, called them 'smallest quanta.' Today they are called 'atoms', in honour of a famous argument of the ancient Greeks. Indeed, 2500 years ago, the Greeks asked the following question. If motion and matter are conserved, how can change and transformation exist? The philosophical school of Leucippus and Democritus of Abdera** studied two

* The human ear can detect pressure variations at least as small as 20 μPa.

check the greek
check the greek

** Leucippus of Elea (Λευκιππος) (*c.* 490 to *c.* 430 BCE), Greek philosopher; Elea was a small town south of Naples. It lies in Italy, but used to belong to the Magna Graecia. Democritus (Δεμοκριτος) of Abdera (*c.* 460 to *c.* 356 or 370 BCE), also a Greek philosopher, was arguably the greatest philosopher who ever lived. Together with his teacher Leucippus, he was the founder of the atomic theory; Democritus was a much admired thinker, and a contemporary of Socrates. The vain Plato never even mentions him, as Democritus

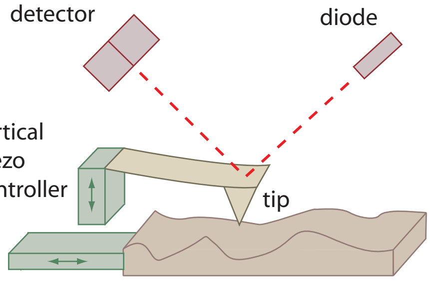


Figure 106 The principle, and a simple realization, of an atomic force microscope

particular observations in special detail. They noted that salt dissolves in water. They also noted that fish can swim in water. In the first case, the volume of water does not increase when the salt is dissolved. In the second case, when fish advance, they must push water aside. They deduced that there is only one possible explanation that satisfies observations and also reconciles conservation and transformation: nature is made of void and of small, hard, indivisible and conserved particles.* In this way any example of motion, change or transformation is due to rearrangements of these particles; change and conservation are reconciled.

In short, since matter is hard, has a shape and is divisible, Leucippus and Democritus imagined it as being made of atoms. Atoms are particles which are hard, have a shape, but are indivisible. In other words, the Greeks imagined nature as a big Lego set. Lego pieces are first of all hard or *impenetrable*, i.e. repulsive at very small distances. They are *attractive* at small distances: they remain stuck together. Finally, they have *no interaction* at large distances. Atoms behave in the same way. (Actually, what the Greeks called 'atoms' partly corresponds to what today we call 'molecules'. The latter term was invented by Amadeo Avogadro in 1811 in order to clarify the distinction. But we can forget this detail for the moment.)

Since atoms are so small, it took many years before all scientists were convinced by the experiments showing their existence. In the nineteenth century, the idea of atoms was Page 220 beautifully verified by the discovery of the 'laws' of chemistry and those of gas behaviour. Later on, the noise effects were discovered.

Nowadays, with advances in technology, single atoms can be seen, photographed, holo- Ref. 162, Ref. 163 grammed, counted, touched, moved, lifted, levitated, and thrown around. And indeed, like everyday matter, atoms have mass, size, shape and colour. Single atoms have even Ref. 164 been used as lamps and lasers.

was a danger to his own fame. Democritus wrote many books which have been lost; they were not copied during the Middle Ages because of his scientific and rational world view, which was felt to be a danger by religious zealots who had the monopoly on the copying industry.

* The story is told by Lucrece, or Titus Lucretius Carus, in his famous text *De natura rerum*, around 50 BCE. Especially if we imagine particles as little balls, we cannot avoid calling this a typically male idea. (What Challenge 413 d would be the female approach?)

no image yet

Figure 107 The atoms on the surface of a silicon crystal mapped with an atomic force microscope

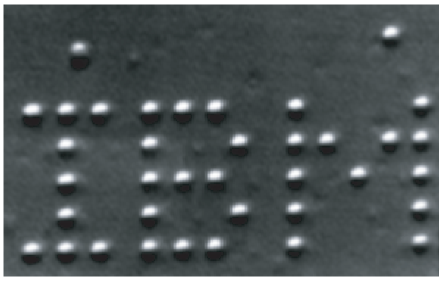


Figure 108 The result of moving helium atoms on a metallic surface (© IBM)

Modern researchers in several fields have fun playing with atoms in the same way that children play with Lego. Maybe the most beautiful demonstration of these possibilities is provided by the many applications of the atomic force microscope. If you ever have the opportunity to see one, do not miss it!* It is a simple device which follows the surface of an object with an atomically sharp needle; such needles, usually of tungsten, are easily Ref. 165 manufactured with a simple etching method. The changes in the height of the needle along its path over the surface are recorded with the help of a deflected light ray. With a little care, the atoms of the object can be felt and made visible on a computer screen. With special types of such microscopes, the needle can be used to move atoms one by one to specified places on the surface. It is also possible to scan a surface, pick up a given atom Ref. 166 and throw it towards a mass spectrometer to determine what sort of atom it is.

Incidentally, the construction of atomic force microscopes is only a small improvement on what nature is building already by the millions; when we use our ears to listen, we are actually detecting changes in eardrum position of about 1 nm. In other words, we all have two 'atomic force microscopes' built into our heads.

In summary, matter is not scale invariant: in particular, it is neither smooth nor fractal. Matter is made of atoms. Different types of atoms, as well as their various combinations, produce different types of substances. Pictures from atomic force microscopes show that the size and arrangement of atoms produce the *shape* and the *extension* of objects, confirming the Lego model of matter.* As a result, the description of the motion of extended objects can be reduced to the description of the motion of their atoms. Atomic motion will be a major theme in the following pages. One of its consequences is especially im-

* A cheap version costs only a few thousand Euros, and will allow you to study the difference between a silicon wafer – crystalline – a flour wafer – granular-amorphous – and consecrated wafer.

* Studying matter in even more detail yields the now well-known idea that matter, at higher and higher magnifications, is made of molecules, atoms, nuclei, protons and neutrons, and finally, quarks. Atoms also contain electrons. A final type of matter, neutrinos, is observed coming from the Sun and from certain types of radioactive materials. Even though the fundamental bricks have become smaller with time, the basic idea remains: matter is made of smallest entities, nowadays called elementary particles. In the second part of our Page 857 mountain ascent we will explore this idea in detail. Appendix C lists the measured properties of all known elementary particles.

portant: heat.

Curiosities and fun challenges about fluids and solids

Before we continue, a few puzzles are due. They indicate the range of phenomena that the motion of extended bodies encompasses.

▪ You are in a boat on a pond with a stone, a bucket of water and a piece of wood. What happens to the water level of the pond after you throw the stone in it? After you throw the water into the pond? After you throw the piece of wood? Challenge 414 n

▪ What is the maximum length of a vertically hanging wire? Could a wire be lowered from a suspended geostationary satellite down to the Earth? This would mean we could realize a space 'elevator'. How long would the cable have to be? How heavy would it be? How would you build such a system? What dangers would it face? Challenge 415 n

▪ Matter is made of atoms. Over the centuries the stubborn resistance of many people to this idea has lead to the loss of many treasures. For over a thousand years, people thought that genuine pearls could be distinguished from false ones by hitting them with a hammer: only false pearls would break. Unfortunately, *all* pearls break. As a result, over time all the most beautiful pearls in the world have been smashed to pieces.

▪ Put a rubber air balloon over the end of a bottle and let it hang inside the bottle. How much can you blow up the balloon inside the bottle? Challenge 416 e

▪ Put a small paper ball into the neck of a horizontal bottle and try to blow it into the bottle. The paper will fly *towards* you. Why? Challenge 417 e

▪ It is possible to blow an egg from one egg-cup to a second one just behind it. Can you to perform this trick? Challenge 418 e

▪ In the seventeenth century, engineers who needed to pump water faced a challenge. To pump water from mine shafts to the surface, no water pump managed more than 10 m of height difference. For twice that height, one always needed two pumps in series, connected by an intermediate reservoir. Why? How then do trees manage to pump water upwards for larger heights? Challenge 419 ny

▪ Comic books have difficulties with the concept of atoms. Could Asterix really throw Romans into the air using his fist? Are Lucky Luke's precise revolver shots possible? Can Spiderman's silk support him in his swings from building to building? Can the Roadrunner stop running in three steps? Can the Sun be made to stop in the sky by command? Can space-ships hover using fuel? Take any comic-book hero and ask yourself whether matter made of atoms would allow him the feats he seems capable of. You will find that most cartoons are comic precisely because they assume that matter is not made of atoms, but continuous! In a sense, atoms make life a serious adventure. Challenge 420 e

▪ When hydrogen and oxygen are combined to form water, the amount of hydrogen needed is exactly twice the amount of oxygen, if no gas is to be left over after the reaction. How does this observation confirm the existence of atoms? Challenge 421 n

▪ How are alcohol-filled chocolate pralines made? Note that the alcohol is not injected into them afterwards, because there would be no way to keep the result tight enough. Challenge 422 n

▪ How often can a stone jump when it is thrown over the surface of water? The present world record was achieved in 1992 when a palm-sized, triangular and flat stone was thrown with a speed of about 12 m/s and a rotation speed of about 14 revolutions per second along a river, covering about 100 m with an astonishing 38 jumps. (The sequence Ref. 167

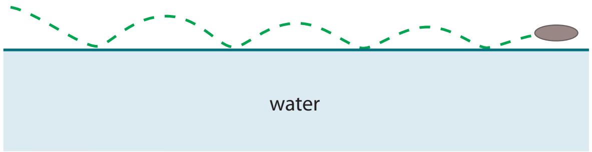


Figure 109 What is your personal stone-skipping record?

was filmed with a video recorder from a bridge.) What would be necessary to increase the number of jumps? Can you build a machine that is a better thrower than yourself?

Challenge 423 r

▪ The biggest component of air is nitrogen (about 78 %). The second biggest component is oxygen (about 21 %). What is the third biggest one?

Challenge 424 n

▪ Water can flow uphill: Heron's fountain shows this most clearly. Heron of Alexandria (*c.* 10 to *c.* 70) described it 2000 years ago; it is easily built at home, using some plastic bottles and a little tubing. How does it work?

Challenge 425 n

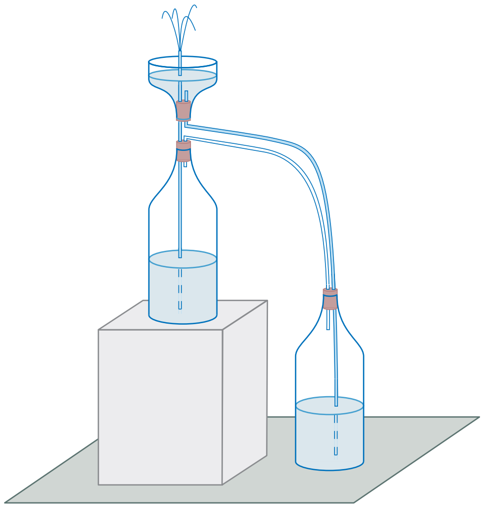


Figure 110 Heron's fountain

▪ A light bulb is placed, underwater, in a stable steel cylinder with a diameter of 16 cm. A Fiat Cinquecento (500 kg) is placed on a piston pushing onto the water surface. Will the bulb resist?

Challenge 426 n

▪ What is the most dense gas? The most dense vapour?

Challenge 427 ny

▪ Every year, the Institute of Maritime Systems of the University of Rostock organizes a contest. The challenge is to build a paper boat with the highest carrying capacity. The paper boat must weigh at most 10 g; the carrying capacity is measured by pouring lead small shot onto it, until the boat sinks. The 2002 record stands at 2.6 kg. Can you achieve this value? (For more information, see the http://www.paperboat.de website.)

Challenge 428 e

▪ A modern version of an old question – already posed by Daniel Colladon (1802–1893) – is the following. A ship of mass m in a river is pulled by horses walking along the riverbank attached by ropes. If the river is of superfluid helium, meaning that there is no friction between ship and river, what energy is necessary to pull the ship upstream along the river until a height h has been gained?

Challenge 429 n

▪ The Swiss professor Auguste Piccard (1884–1962) was a famous explorer of the stratosphere. He reached a height of 16 km in his *aerostat*. Inside the airtight cabin hanging under his balloon, he had normal air pressure. However, he needed to introduce several ropes attached at the balloon into the cabin, in order to be able to pull them, as they controlled his balloon. How did he get the ropes into the cabin while preventing air from leaving the cabin?

Challenge 430 n

▪ A human cannot breathe at any depth under water, even if he has a tube going to the surface. At a few metres of depth, trying to do so is inevitably fatal! Even at a depth of 60 cm only, the human body can only breathe in this way for a few minutes. Why?

Challenge 431 n

▪ A human in air falls with a limiting speed of about 180 km/h, depending on clothing. How long does it take to fall from 3000 m to 200 m? Challenge 432 ny

▪ Several humans have survived free falls from aeroplanes for a thousand metres or more, even though they had no parachute. How was this possible? Challenge 433 n

▪ Liquid pressure depends on height. If the average human blood pressure at the height of the heart is 13.3 kPa, can you guess what it is inside the feet when standing? Challenge 434 n

▪ The human heart pumps blood at a rate of about 0.1 l/s. A capillary has the diameter of a red blood cell, around 7 μm, and in it the blood moves at a speed of half a millimetre per second. How many capillaries are there in a human? Challenge 435 n

▪ A few drops of tea usually flow along the underside of the spout of a teapot (or fall onto the table). This phenomenon has even been simulated using supercomputer simulations of the motion of liquids, by Kistler and Scriven, using the Navier–Stokes equations. Ref. 168 Teapots are still shedding drops, though.

▪ The best giant soap bubbles can be made by mixing 1.5 l of water, 200 ml of corn syrup and 450 ml of washing-up liquid. Mix everything together and then let it rest for four hours. You can then make the largest bubbles by dipping a metal ring of up to 100 mm diameter into the mixture. But why do soap bubbles burst? Challenge 436 n

▪ Can humans start earthquakes? What would happen if all the 1000 million Indians were to jump at the same time from the kitchen table to the floor? Challenge 437 n

In fact, several strong earthquakes *have* been triggered by humans. This has happened when water dams have been filled, or when water has been injected into drilling holes. It has been suggested that the extraction of deep underground water also causes earthquakes. If this is confirmed, a sizeable proportion of all earthquakes could be human-triggered.

Challenge 438 n ▪ How can a tip of a stalactite be distinguished from a tip of a stalagmite? Does the difference exist also for icicles?

▪ A drop of water that falls into a pan containing hot oil dances on the surface for a considerable time, if the oil is above 220 °C. Cooks test the temperature of oil in this way. Why does this so-called Leidenfrost effect* take place? Challenge 439 ny

▪ How much more weight would your bathroom scales show if you stood on them in a vacuum? Challenge 440 n

Challenge 441 n ▪ Why don't air molecules fall towards the bottom of the container and stay there?

Challenge 442 n ▪ Which of the two water funnels in Figure 111 is emptied more rapidly? Apply energy conservation to the fluid's motion (also called Bernoulli's 'law') to find the answer.

▪ As we have seen, fast flow generates an underpressure. How do fish prevent their eyes from popping when they swim rapidly? Challenge 443 n

▪ Golf balls have dimples for the same reasons that tennis balls are hairy and that shark and dolphin skin is not flat: deviations from flatness reduce the flow resistance because many small eddies produce less friction than a few large ones. Why? Challenge 444 ny

▪ One of the most complex extended bodies is the human body. In modern simulations of the behaviour of humans in car accidents, the most advanced models include ribs, vertebrae, all other bones and the various organs. For each part, its specific deformation properties are taken into account. With such models and simulations, the protection of passengers and drivers in cars can be optimized.

* It is named after Johann Gottlieb Leidenfrost (1715–1794), German physician.

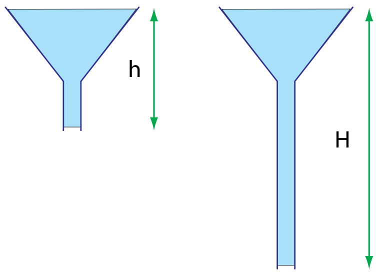


Figure 111 Which funnel is faster?

▪ Glass is a solid. Nevertheless, many textbooks say that glass is a liquid. This error has been propagated for about a hundred years, probably originating from a mistranslation of a sentence in a German textbook published in 1933 by Gustav Tamman, *Der Glaszustand*. Can you give at least three reasons why glass is a solid and not a liquid? Challenge 445 n

▪ The recognized record height reached by a helicopter is 12 442 m above sea level, though 12 954 m has also been claimed. (The first height was reached in 1972, the second in 2002, both by French pilots in French helicopters.) Why, then, do people still continue to use their legs in order to reach the top of Mount Sagarmatha, the highest mountain in the world? Challenge 446 n

▪ A loosely knotted sewing thread lies on the surface of a bowl filled with water. Putting a bit of washing-up liquid into the area surrounded by the thread makes it immediately become circular. Why? Challenge 447 e

▪ The deepest hole ever drilled into the Earth is 12 km deep. In 2003, somebody proposed to enlarge such a hole and then to pour millions of tons of liquid iron into it. He claims that the iron would sink towards the centre of the Earth. If a measurement device communication were dropped into the iron, it could send its observations to the surface using sound waves. Can you give some reasons why this would not work? Challenge 448 n

▪ How can you put a handkerchief under water using a glass, while keeping it dry? Challenge 449 n

▪ Are you able to blow a ping pong ball out of a funnel? What happens if you blow through a funnel towards a burning candle?

▪ The economic power of a nation has long been associated with its capacity to produce high-quality steel. Indeed, the Industrial Revolution started with the mass production of steel. Every scientist should know the basics facts about steel. *Steel* is a combination of iron and carbon to which other elements, mostly metals, may be added as well. One can distinguish three main types of steel, depending on the crystalline structure. *Ferritic steels* have a body-centred cubic structure, *austenitic steels* have a face-centred cubic structure, and *martensitic steels* have a body-centred tetragonal structure. Table 26 gives further details.

▪ A simple phenomenon which requires a complex explanation is the cracking of a whip. Since the experimental work of Peter Krehl it has been known that the whip cracks when the tip reaches a velocity of *twice* the speed of sound. Can you imagine why? Ref. 169 Challenge 450 ny

▪ The fall of a leaf, with its complex path, is still a topic of investigation. We are far from being able to predict the time a leaf will take to reach the ground; the motion of the air around a leaf is not easy to describe. On of the simplest phenomena of hydrodynamics remains one of its most difficult problems.

▪ Fluids exhibit many interesting effects. Soap bubbles in air are made of a thin spherical film of liquid with air on both sides. In 1932, anti-bubbles, thin spherical films of air with liquid on both sides, were first observed. In 2004, the Belgian physicist Stéphane Dorbolo and his team showed that it is possible to produce them in simple experiments, and in particular, in Belgian beer. Ref. 170

Table 26 Steel types, properties and uses

Ferritic steel	Austenitic steel	Martensitic steel
'usual' steel	'soft' steel	hardened steel, brittle
body centred cubic (bcc)	face centred cubic (fcc)	body centred tetragonal (bct)
iron and carbon	iron, chromium, nickel, manganese, carbon	carbon steel and alloys
Examples		
construction steel	most stainless (18/8 Cr/Ni) steels	knife edges
car sheet steel	kitchenware	drill surfaces
ship steel	food industry	spring steel, crankshafts
12 % Cr stainless ferrite	Cr/V steels for nuclear reactors	
Properties		
phases described by the iron-carbon phase diagram	phases described by the Schaeffler diagram	phases described by the iron-carbon diagram and the TTT (time–temperature transformation) diagram
in equilibrium at RT	some alloys in equilibrium at RT	not in equilibrium at RT, but stable
mechanical properties and grain size depend on heat treatment	mechanical properties and grain size depend on thermo-mechanical pre-treatment	mechanical properties and grain size strongly depend on heat treatment
hardened by reducing grain size, by forging, by increasing carbon content or by nitration	hardened by cold working only	hard anyway – made by laser irradiation, induction heating, etc.
grains of ferrite and paerlite, with cementite (Fe_3C)	grains of austenite	grains of martensite
ferromagnetic	not magnetic or weakly magnetic	ferromagnetic

What can move in nature?

Before we continue to the next way to describe motion globally, we will have a look at the possibilities of motion in everyday life. One overview is given in Table 27. The domains that belong to everyday life – motion of fluids, of matter, of matter types, of heat, of light and of charge – are the domains of continuum physics.

Within continuum physics, there are three domains we have not yet studied: the motion of charge and light, called electrodynamics, the motion of heat, called thermodynamics, and the motion of the vacuum. Once we have explored these domains, we will have completed the first step of our description of motion: continuum physics. In continuum physics, motion and moving entities are described with continuous quantities that can take any value, including arbitrarily small or arbitrarily large values.

Table 27 Extensive quantities in nature, i.e. quantities that *flow* and *accumulate*

Domain	Extensive quantity (energy carrier)	Current (flow intensity)	Intensive quantity (driving strength)	Energy flow (power)	Resistance to transport (intensity of entropy generation)
Rivers	mass m	mass flow m/t	height difference gh	$P = gh\, m/t$	$R_\mathrm{m} = ght/m$ $[\mathrm{m^2/s\,kg}]$
Gases	volume V	volume flow V/t	pressure p	$P = pV/t$	$R_\mathrm{V} = pt/V$ $[\mathrm{kg/s\,m^5}]$
Mechanics	momentum $\mathbf{p}$	force $\mathbf{F} = \mathrm{d}\mathbf{p}/\mathrm{d}t$	velocity $\mathbf{v}$	$P = \mathbf{v}\,\mathbf{F}$	$R_\mathrm{p} = t/m$ $[\mathrm{s/kg}]$
	angular momentum $\mathbf{L}$	torque $\mathbf{M} = \mathrm{d}\mathbf{L}/\mathrm{d}t$	angular velocity $\boldsymbol{\omega}$	$P = \boldsymbol{\omega}\,\mathbf{M}$	$R_\mathrm{L} = t/mr^2$ $[\mathrm{s/kg\,m^2}]$
Chemistry	amount of substance n	substance flow $I_n = \mathrm{d}n/\mathrm{d}t$	chemical potential μ	$P = \mu\, I_n$	$R_n = \mu t/n$ $[\mathrm{Js/mol^2}]$
Thermo-dynamics	entropy S	entropy flow $I_S = \mathrm{d}S/\mathrm{d}t$	temperature T	$P = T\, I_S$	$R_S = Tt/S$ $[\mathrm{K^2/W}]$
Light	like all massless radiation, it can flow but cannot accumulate				
Electricity	charge q	electrical current $I = \mathrm{d}q/\mathrm{d}t$	electrical potential U	$P = U\, I$	$R = U/I$ $[\Omega]$
Magnetism	no accumulable magnetic sources are found in nature				
Nuclear physics	extensive quantities exist, but do not appear in everyday life				
Gravitation	empty space can move and flow, but the motion is not observed in everyday life				

But nature is *not* continuous. We have already seen that matter cannot be indefinitely divided into ever-smaller entities. In fact, we will discover that there are precise experiments that provide limits to the observed values for *every* domain of continuum physics. There is a limit to mass, to speed, to angular momentum, to force, to entropy and to change of charge. The consequences of these discoveries form the second step in our description of motion: quantum theory and relativity. Quantum theory is based on lower limits; relativity is based on upper limits. The third and last step of our description of motion will be formed by the unification of quantum theory and general relativity.

Every domain of physics, regardless of which one of the above steps it belongs to, describes change in terms two quantities: energy, and an extensive quantity characteristic of the domain. An observable quantity is called *extensive* if it increases with system size. Table 27 provides an overview. The intensive and extensive quantities corresponding to

Ref. 188

what in everyday language is called 'heat' are *temperature* and *entropy*.

Why are objects warm?

We continue our short stroll through the field of global descriptions of motion with an overview of heat and the main concepts associated with it. For our purposes we only need to know the basic facts about heat. The main points that are taught in school are almost sufficient. Ref. 171

Macroscopic bodies, i.e. bodies made of many atoms, have temperature. The temperature of a macroscopic body is an aspect of its state. It is observed that any two bodies in contact tend towards the same temperature: temperature is contagious. In other words, temperature describes an equilibrium situation. The existence and contagiousness of temperature is often called the *zeroth principle of thermodynamics*. Heating is the increase of temperature.

How is temperature measured? The eighteenth century produced the clearest answer: temperature is best defined and measured by the *expansion of gases*. For the simplest, so-called *ideal* gases, the product of pressure p and volume V is proportional to temperature:

$$pV \sim T \ . \tag{88}$$

The proportionality constant is fixed by the *amount* of gas used. (More about it shortly.) The ideal gas relation allows us to determine temperature by measuring pressure and volume. This is the way (absolute) temperature has been defined and measured for about a century. To define the *unit* of temperature, one only has to fix the amount of gas used. It is customary to fix the amount of gas at 1 mol; for oxygen this is 32 g. The proportionality constant, called the *ideal gas constant* R, is defined to be $R = 8.3145\,\mathrm{J/mol\,K}$. This number has been chosen in order to yield the best approximation to the independently defined Celsius temperature scale. Fixing the ideal gas constant in this way defines 1 K, or one Kelvin, as the unit of temperature. In simple terms, a temperature increase of one Kelvin is defined as the temperature increase that makes the volume of an ideal gas increase – keeping the pressure fixed – by a fraction of 1/273.15 or 0.3661 %. Ref. 175 Page 1060 Challenge 451 ny

In general, if one needs to determine the temperature of an object, one takes a mole of gas, puts it in contact with the object, waits a while, and then measures the pressure and the volume of the gas. The ideal gas relation (88) then gives the temperature. Most importantly, the ideal gas relation shows that there is a lowest temperature in nature, namely that temperature at which an ideal gas would have a vanishing volume. That would happen at $T = 0\,\mathrm{K}$, i.e. at $-273.15\,°\mathrm{C}$. Obviously, other effects, like the volume of the atoms themselves, prevent the volume of the gas from ever reaching zero. The *third principle of thermodynamics* provides another reason why this is impossible.

The temperature achieved by a civilization can be used as a measure of its technological achievements. One can define the Bronze Age (1.1 kK, 3500 BCE) , the Iron Age (1.8 kK, 1000 BCE), the Electric Age (3 kK from *c.* 1880) and the Atomic Age (several MK, from 1944) in this way. Taking into account also the quest for lower temperatures, one can define the Quantum Age (4 K, starting 1908). Ref. 176

Heating implies flow of energy. For example, friction heats up and slows down mov-

Table 28 Some temperature values

Observation	Temperature
Lowest, but unattainable, temperature	0 K = −273.15°C
In the context of lasers, it sometimes makes sense to talk about negative temperature.	
Temperature a perfect vacuum would have at Earth's surface Page 812	40 zK
Sodium gas in certain laboratory experiments – coldest matter system achieved by man and possibly in the universe	0.45 nK
Temperature of neutrino background in the universe	*c.* 2 K
Temperature of photon gas background (or background radiation) in the universe	2.7 K
Liquid helium	4.2 K
Oxygen triple point	54.3584 K
Liquid nitrogen	77 K
Coldest weather ever measured (Antarctic)	185 K = −88°C
Freezing point of water at standard pressure	273.15 K = 0.00°C
Triple point of water	273.16 K = 0.01°C
Average temperature of the Earth's surface	287.2 K
Interior of human body	310.0 ± 0.5 K = 36.8 ± 0.5°C
Hottest weather measured	331 K = 58°C
Boiling point of water at standard pressure	373.13 K or 99.975°C
Liquid bronze	*c.* 1100 K
Liquid, pure iron	1810 K
Freezing point of gold	1337.33 K
Light bulb filament	2.9 kK
Earth's centre	4 kK
Sun's surface	5.8 kK
Air in lightning bolt	30 kK
Hottest star's surface (centre of NGC 2240)	250 kK
Space between Earth and Moon (no typo)	up to 1 MK
Sun's centre	20 MK
Inside the JET fusion tokamak	100 MK
Centre of hottest stars	1 GK
Maximum temperature of systems without electron–positron pair generation	ca. 6 GK
Universe when it was 1 s old	100 GK
Hagedorn temperature	1.9 TK
Heavy ion collisions – highest man-made value	up to 3.6 TK
Planck temperature – nature's upper temperature limit	10^{32} K

ing bodies. In the old days, the 'creation' of heat by friction was even tested experimentally. It was shown that heat could be generated from friction, just by continuous rubbing, without any limit; this 'creation' implies that heat is not a material fluid extracted from the body – which in this case would be consumed after a certain time – but something else. Indeed, today we know that heat, even though it behaves in some ways like a fluid, is due to disordered motion of particles. The conclusion of these studies is simple. Friction is the transformation of mechanical energy into *thermal energy*.

To heat 1 kg of water by 1 K by friction, 4.2 kJ of mechanical energy must be transformed through friction. The first to measure this quantity with precision was, in 1842, the German physician Julius Robert Mayer (1814–1878). He regarded his experiment as proof of the conservation of energy; indeed, he was the first person to state energy conservation! It is something of an embarrassment to modern physics that a medical doctor was the first to show the conservation of energy, and furthermore, that he was ridiculed by most physicists of his time. Worse, conservation of energy was accepted only when it was repeated many years later by two authorities: Hermann von Helmholtz – himself also a physician turned physicist – and William Thomson, who also cited similar, but later experiments by James Joule.* All of them acknowledged Mayer's priority. Publicity by William Thomson eventually led to the naming of the unit of energy after Joule.

In short, the sum of mechanical energy and thermal energy is constant. This is usually called the *first principle of thermodynamics*. Equivalently, it is impossible to produce mechanical energy without paying for it with some other form of energy. This is an important statement, because among others it means that humanity will stop living one day. Indeed, we live mostly on energy from the Sun; since the Sun is of finite size, its energy content will eventually be consumed. Can you estimate when this will happen? Challenge 452 n

There is also a second (and the mentioned third) principle of thermodynamics, which will be presented later on. The study of these topics is called *thermostatics* if the systems concerned are at equilibrium, and *thermodynamics* if they are not. In the latter case, we distinguish situations *near* equilibrium, when equilibrium concepts such as temperature can still be used, from situations *far* from equilibrium, such as self-organization, where such concepts often cannot be applied. Page 232

Does it make sense to distinguish between thermal energy and heat? It does. Many older texts use the term 'heat' to mean the same as thermal energy. However, this is confusing; in this text, 'heat' is used, in accordance with modern approaches, as the everyday term for entropy. Both thermal energy and heat flow from one body to another, and both accumulate. Both have no measurable mass.** Both the amount of thermal energy and the amount of heat inside a body increase with increasing temperature. The precise relation will be given shortly. But heat has many other interesting properties and stories to tell. Of these, two are particularly important: first, heat is due to particles; and secondly, heat is at the heart of the difference between past and future. These two stories are intertwined.

* Hermann von Helmholtz (b. 1821 Potsdam, d. 1894 Berlin), important Prussian scientist. William Thomson (later William Kelvin) (1824–1907), important Irish physicist. James Prescott Joule (1818–1889), English physicist. Joule is pronounced so that it rhymes with 'cool', as his descendants like to stress. (The pronunciation of the name 'Joule' varies from family to family.)

** This might change in future, when mass measurements improve in precision, thus allowing the detection of relativistic effects. In this case, temperature increase may be detected through its related mass increase. However, such changes are noticeable only with twelve or more digits of precision in mass measurements. Page 286

Table 29 Some measured entropy values

Process/System	Entropy value
Melting of 1 kg of ice	1.21 kJ/K kg = 21.99 J/K mol
Water under standard conditions	70.1 J/K mol
Boiling of 1 kg of liquid water at 101.3 kPa	6.03 kJ/K= 110 J/K mol
Iron under standard conditions	27.2 J/K mol
Oxygen under standard conditions	161.1 J/K mol

Entropy

– It's irreversible.
– Like my raincoat!

Mel Brooks, *Spaceballs*, 1987

Every domain of physics describes change in terms of two quantities: energy, and an extensive quantity characteristic of the domain. Even though heat is related to energy, the quantity physicists usually call heat is *not* an extensive quantity. Worse, what physicists call heat is not the same as what we call heat in our everyday speech. The extensive quantity corresponding to what we call 'heat' in everyday speech is called *entropy*.* Entropy describes heat in the same way as momentum describes motion. When two objects differing in temperature are brought into contact, an entropy flow takes place between them, like the flow of momentum that take place when two objects of different speeds collide. Let us define the concept of entropy more precisely and explore its properties in some more detail.

Ref. 188

Entropy measures the degree to which energy is *mixed up* inside a system, that is, the degree to which energy is spread or shared among the components of a system. Therefore, entropy adds up when identical systems are composed into one. When two litre bottles of water at the same temperature are poured together, the entropy of the water adds up.

Like any other extensive quantity, entropy can be accumulated in a body; it can flow into or out of bodies. When water is transformed into steam, the entropy added into the water is indeed contained in the steam. In short, entropy is what is called 'heat' in everyday speech.

In contrast to several other important extensive quantities, entropy is not conserved. The sharing of energy in a system can be increased, for example by heating it. However, entropy is 'half conserved': in closed systems, entropy does not decrease; mixing cannot be undone. What is called equilibrium is simply the result of the highest possible mixing. In short, the entropy in a closed system increases until it reaches the maximum possible value.

When a piece of rock is detached from a mountain, it falls, tumbles into the valley, heating up a bit, and eventually stops. The opposite process, whereby a rock cools and tumbles upwards, is never observed. Why? The opposite motion does not contradict any rule or pattern about motion that we have deduced so far.

Challenge 453 ny

* The term 'entropy' was invented by the German physicist Rudolph Clausius (1822–1888) in 1865. He

Rocks never fall upwards because mountains, valleys and rocks are made of many particles. Motions of many-particle systems, especially in the domain of thermostatics, are called *processes*. Central to thermostatics is the distinction between *reversible* processes, such as the flight of a thrown stone, and *irreversible* processes, such as the aforementioned tumbling rock. Irreversible processes are all those processes in which friction and its generalizations play a role. They are those which increase the sharing or mixing of energy. They are important: if there were no friction, shirt buttons and shoelaces would not stay fastened, we could not walk or run, coffee machines would not make coffee, and maybe most importantly of all, we would have no memory.

Ref. 189
Page 751

Irreversible processes, in the sense in which the term is used in thermostatics, transform macroscopic motion into the disorganized motion of all the small microscopic components involved: they increase the sharing and mixing of energy. Irreversible processes are therefore not *strictly* irreversible – but their reversal is extremely improbable. We can say that entropy measures the 'amount of irreversibility': it measures the degree of mixing or decay that a collective motion has undergone.

Entropy is not conserved. Entropy – 'heat' – can appear out of nowhere, since energy sharing or mixing can happen by itself. For example, when two different liquids of the same temperature are mixed – such as water and sulphuric acid – the final temperature of the mix can differ. Similarly, when electrical current flows through material at room temperature, the system can heat up or cool down, depending on the material.

The *second principle of thermodynamics* states that 'entropy ain't what it used to be.' More precisely, *the entropy in a closed system tends towards its maximum*. Here, a *closed system* is a system that does not exchange energy or matter with its environment. Can you think of an example?

Challenge 454 ny

Entropy never decreases. Everyday life shows that in a closed system, the disorder increases with time, until it reaches some maximum. To reduce disorder, we need effort, i.e. work and energy. In other words, in order to reduce the disorder in a system, we need to connect the system to an energy source in some clever way. Refrigerators need electrical current precisely for this reason.

Because entropy never decreases, *white colour does not last*. Whenever disorder increases, the colour white becomes 'dirty', usually grey or brown. Perhaps for this reason white objects, such as white clothes, white houses and white underwear, are valued in our society. White objects defy decay.

Entropy allows to define the concept of *equilibrium* more precisely as the state of maximum entropy, or maximum energy sharing.

Flow of entropy

We know from daily experience that transport of an extensive quantity always involves friction. Friction implies generation of entropy. In particular, the flow of entropy itself produces additional entropy. For example, when a house is heated, entropy is produced in the wall. Heating means to keep a temperature difference ΔT between the interior and

formed it from the Greek ἐν 'in' and τρόπος 'direction', to make it sound similar to 'energy'. It has always had the meaning given here.

the exterior of the house. The heat flow J traversing a square meter of wall is given by

$$J = \kappa \Delta T = \kappa (T_i - T_e) \tag{89}$$

where κ is a constant characterizing the ability of the wall to conduct heat. While conducting heat, the wall also *produces* entropy. The entropy production σ is proportional to the difference between the interior and the exterior entropy flows. In other words, one has

$$\sigma = \frac{J}{T_e} - \frac{J}{T_i} = \kappa \frac{(T_i - T_e)^2}{T_i T_e} \; . \tag{90}$$

Note that we have assumed in this calculation that everything is near equilibrium in each slice parallel to the wall, a reasonable assumption in everyday life. A typical case of a good wall has $\kappa = 1\,\mathrm{W/m^2K}$ in the temperature range between 273 K and 293 K. With this value, one gets an entropy production of

$$\sigma = 5 \cdot 10^{-3}\,\mathrm{W/m^2K} \; . \tag{91}$$

Challenge 455 ny

Can you compare the amount of entropy that is produced in the flow with the amount that is transported? In comparison, a good goose-feather duvet has $\kappa = 1.5\,\mathrm{W/m^2K}$, which in shops is also called 15 tog.*

There are two other ways, apart from heat conduction, to transport entropy: *convection*, used for heating houses, and *radiation*, which is possible also through empty space. For example, the Earth radiates about $1.2\,\mathrm{W/m^2K}$ into space, in total thus about 0.51 PW/K. The entropy is (almost) the same that the Earth receives from the Sun. If more entropy had to be radiated away than received, the temperature of the surface of the Earth would have to increase. This is called the *greenhouse effect*. (It is also called *global warming*.) Let's hope that it remains small in the near future.

Do isolated systems exist?

In all our discussions so far, we have assumed that we can distinguish the system under investigation from its environment. But do such *isolated* or *closed* systems, i.e. systems not interacting with their environment, actually exist? Probably our own human condition was the original model for the concept: we do experience having the possibility to act independently of our environment. An isolated system may be simply defined as a system not exchanging any energy or matter with its environment. For many centuries, scientists saw no reason to question this definition.

The concept of an isolated system had to be refined somewhat with the advent of quantum mechanics. Nevertheless, the concept provides useful and precise descriptions

* That unit is not as bad as the official (not a joke) BthU · h/sqft/cm/°F used in some remote provinces of our galaxy.

The insulation power of materials is usually measured by the constant $\lambda = \kappa d$ which is independent of the thickness d of the insulating layer. Values in nature range from about 2000 W/K m for diamond, which is the best conductor of all, down to between 0.1 W/K m and 0.2 W/K m for wood, between 0.015 W/K m and 0.05 W/K m for wools, cork and foams, and the small value of $5 \cdot 10^{-3}$ W/K m for krypton gas.

of nature also in that domain. Only in the third part of our walk will the situation change drastically. There, the investigation of whether the universe is an isolated system will lead to surprising results. (What do you think?)* We'll take the first steps towards the answer shortly.

Challenge 456 n

Why do balloons take up space? – The end of continuity

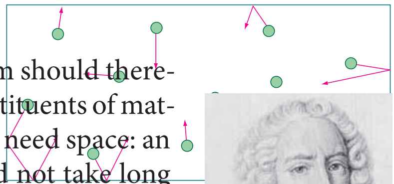
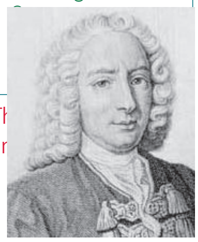


Figure 112 Th
statistical mechan



Daniel Bernoulli

Heat properties are material-dependent. Studying them should therefore enable us to understand something about the constituents of matter. Now, the simplest materials of all are gases.* Gases need space: an amount of gas has pressure and volume. Indeed, it did not take long to show that gases *could not* be continuous. One of the first scientists to think about gases as made up of atoms was Daniel Bernoulli.** Bernoulli reasoned that if atoms are small particles, with mass and momentum, he should be able to make quantitative predictions about the behaviour of gases, and check them with experiment. If the particles fly around in a gas, then the *pressure* of a gas in a container is produced by the steady flow of particles hitting the wall. It was then easy to conclude that if the particles are assumed to behave as tiny, hard and perfectly elastic balls, the pressure p, volume V and temperature T must be related by

Challenge 457 ny

$$pV = \frac{3k}{2} N T \qquad (92)$$

where N is the number of particles contained in the gas. (The Boltzmann constant k, one of the fundamental constants of nature, is defined below.) A gas made of particles with such textbook behaviour is called an *ideal gas*. Relation (92) has been confirmed by experiments at room and higher temperatures, for all known gases.

Bernoulli thus derived the gas relation, with a specific prediction for the proportionality constant, from the single assumption that gases are made of small massive constituents. This derivation provides a clear argument for the existence of atoms and for their behaviour as normal, though small objects. (Can you imagine how N might be determined experimentally?)

Challenge 458 ny

* A strange hint: your answer is almost surely wrong.

* By the way, the word *gas* is a modern construct. It was coined by the Brussels alchemist and physician Johan Baptista van Helmont (1579–1644), to sound similar to 'chaos'. It is one of the few words which have been invented by one person and then adopted all over the world.

** Daniel Bernoulli (b. 1700 Bâle, d. 1782 Bâle), important Swiss mathematician and physicist. His father Johann and his uncle Jakob were famous mathematicians, as were his brothers and some of his nephews. Daniel Bernoulli published many mathematical and physical results. In physics, he studied the separation of compound motion into translation and rotation. In 1738 he published the *Hydrodynamique*, in which he deduced all results from a single principle, namely the conservation of energy. The so-called *Bernoulli's principle* states that (and how) the pressure of a fluid decreases when its speed increases. He studied the tides and many complex mechanical problems, and explained the Boyle–Mariotte gas law. For his publications he won the prestigious prize of the French Academy of Sciences – a forerunner of the Nobel prize – ten times.

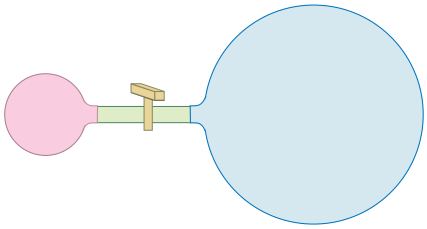


Figure 113 Which balloon wins?

The ideal gas model helps us to answer questions such as the one illustrated in Figure 113. Two *identical* rubber balloons, one filled up to a larger size than the other, are connected via a pipe and a valve. The valve is opened. Which one deflates?

Challenge 459 n

Now you can take up the following challenge: how can you measure the weight of a car or a bicycle with a ruler only?

Challenge 460 n

The picture of gases as being made of hard constituents without any long-distance interactions breaks down at very low temperatures. However, the ideal gas relation (92) can be improved to overcome these limitations by taking into account the deviations due to interactions between atoms or molecules. This approach is now standard practice and allows us to measure temperatures even at extremely low values. The effects observed below 80 K, such as the solidification of air, frictionless transport of electrical current, or frictionless flow of liquids, form a fascinating world of their own, the beautiful domain of low-temperature physics; it will be explored later on.

Ref. 177 Ref. 178 Page 787, page 789

Brownian motion

It is easy to observe, under a microscope, that small particles (such as pollen) in a liquid never come to rest. They seem to follow a random zigzag movement. In 1827, the English botanist Robert Brown (1773–1858) showed with a series of experiments that this observation is independent of the type of particle and of the type of liquid. In other words, Brown had discovered a fundamental noise in nature. Around 1860, this motion was attributed to the molecules of the liquid colliding with the particles. In 1905 and 1906, Marian von Smoluchowski and, independently, Albert Einstein argued that this theory could be tested experimentally, even though at that time nobody was able to observe molecules directly. The test makes use of the specific properties of thermal noise.

Ref. 172

It had already been clear for a long time that if molecules, i.e. indivisible matter particles, really existed, then heat had to be disordered motion of these constituents and temperature had to be the average energy per degree of freedom of the constituents. Bernoulli's model of Figure 112 implies that for monoatomic gases the kinetic energy T_{kin} per particle is given by

Challenge 461 ny

$$T_{\text{kin}} = \frac{3}{2} kT \tag{93}$$

where T is temperature. The so-called *Boltzmann constant* $k = 1.4 \cdot 10^{-23}$ J/K is the standard conversion factor between temperature and energy.* At a room temperature of 293 K,

* The important Austrian physicist Ludwig Boltzmann (b. 1844 Vienna, d. 1906 Duino) is most famous

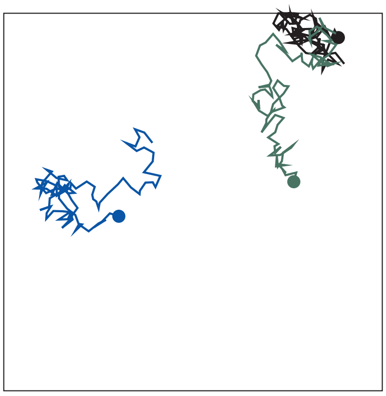


Figure 114 Example paths for particles in Brownian motion and its displacement distribution

the kinetic energy is thus 6 zJ.

Using relation (93) to calculate the speed of air molecules at room temperature yields values of several hundred metres per second. Why then does smoke from a candle take so long to diffuse through a room? Rudolph Clausius (1822–1888) answered this question in the mid-nineteenth century: diffusion is slowed by collisions with air molecules, in the same way as pollen particles collide with molecules in liquids. Challenge 462 ny

At first sight, one could guess that the average distance the pollen particle has moved after n collisions should be zero, because the molecule velocities are random. However, this is wrong, as experiment shows.

An average *square* displacement, written $\langle d^2 \rangle$, is observed for the pollen particle. It cannot be predicted in which direction the particle will move, but it does move. If the distance the particle moves after one collision is l, the average square displacement after n collisions is given, as you should be able to show yourself, by Challenge 463 ny

$$\langle d^2 \rangle = n l^2 \ . \tag{94}$$

For molecules with an average velocity v over time t this gives

$$\langle d^2 \rangle = n l^2 = v l t \ . \tag{95}$$

In other words, the average square displacement increases proportionally with time. Of course, this is only valid if the liquid is made of separate molecules. Repeatedly measuring the position of a particle should give the distribution shown in Figure 114 for the probability that the particle is found at a given distance from the starting point. This is called the *(Gaussian) normal distribution*. In 1908, Jean Perrin* performed extensive experiments in Ref. 173

for his work on thermodynamics, in which he explained all thermodynamic phenomena and observables, including entropy, as results of the behaviour of molecules. Planck named the Boltzmann constant after his investigations. He was one of the most important physicists of the late nineteenth century and stimulated many developments that led to quantum theory. It is said that Boltzmann committed suicide partly because of the resistance of the scientific establishment to his ideas. Nowadays, his work is standard textbook material.

* Jean Perrin (1870–1942), important French physicist, devoted most of his career to the experimental proof

Table 30 Some typical entropy values per particle at *standard* temperature and pressure as multiples of the Boltzmann constant

Material	Entropy per particle
Monoatomic solids	0.3 k to 10 k
Diamond	0.29 k
Graphite	0.68 k
Lead	7.79 k
Monoatomic gases	15-25 k
Helium	15.2 k
Radon	21.2 k
Diatomic gases	15 k to 30 k
Polyatomic solids	10 k to 60 k
Polyatomic liquids	10 k to 80 k
Polyatomic gases	20 k to 60 k
Icosane	112 k

order to test this prediction. He found that equation (95) corresponded completely with observations, thus convincing everybody that Brownian motion is indeed due to collisions with the molecules of the surrounding liquid, as Smoluchowski and Einstein had predicted.* Perrin received the 1926 Nobel Prize for these experiments.

Einstein also showed that the same experiment could be used to determine the number of molecules in a litre of water (or equivalently, the Boltzmann constant k). Can you work out how he did this? Challenge 464 d

Entropy and particles

Once it had become clear that heat and temperature are due to the motion of microscopic particles, people asked what entropy *was* microscopically. The answer can be formulated in various ways. The two most extreme answers are:

- Entropy is the expected number of yes-or-no questions, multiplied by $k \ln 2$, the answers of which would tell us everything about the system, i.e. about its microscopic state.
- Entropy measures the (logarithm of the) number W of possible microscopic states. A given macroscopic state can have many microscopic realizations. The logarithm of this

of the atomic hypothesis and the determination of Avogadro's number; in pursuit of this aim he perfected the use of emulsions, Brownian motion and oil films. His Nobel Prize speech (http://nobelprize.org/physics/laureates/1926/perrin-lecture.html) tells the interesting story of his research. He wrote the influential book *Les atomes* and founded the Centre National de la Recherche Scientifique. He was also the first to speculate, in 1901, that an atom is similar to a small solar system.

* In a delightful piece of research, Pierre Gaspard and his team showed in 1998 that Brownian motion is also chaotic, in the strict physical sense given later on. Ref. 174 Page 236

number, multiplied by the Boltzmann constant k, gives the entropy.*

In short, the higher the entropy, the more microstates are possible. Through either of these definitions, entropy measures the quantity of randomness in a system. In other words, it measures the transformability of energy: higher entropy means lower transformability. Alternatively, entropy measures the *freedom* in the choice of microstate that a system has. High entropy means high freedom of choice for the microstate. For example, when a molecule of glucose (a type of sugar) is produced by photosynthesis, about 40 bits of entropy are released. This means that after the glucose is formed, 40 additional yes-or-no questions must be answered in order to determine the full microscopic state of the system. Physicists often use a macroscopic unit; most systems of interest are large, and thus an entropy of 10^{23} bits is written as 1 J/K.*

Ref. 190 To sum up, entropy is thus a specific measure for the characterization of disorder of thermal systems. Three points are worth making here. First of all, entropy is not *the* measure of disorder, but *one* measure of disorder. It is therefore *not* correct to use entropy as a *synonym* for the concept of disorder, as is often done in the popular literature. Entropy is only defined for systems that have a temperature, in other words, only for systems that are in or near equilibrium. (For systems far from equilibrium, no measure of disorder has been found yet; probably none is possible.) In fact, the use of the term entropy has degenerated so much that sometimes one has to call it *thermodynamic* entropy for clarity.

Secondly, entropy is related to information *only if* information is defined also as $-k \ln W$. To make this point clear, take a book with a mass of one kilogram. At room temperature, its entropy content is about 4 kJ/K. The printed information inside a book, say 500 pages of 40 lines with each containing 80 characters out of 64 possibilities, corresponds to an entropy of $4 \cdot 10^{-17}$ J/K. In short, what is usually called 'information' in everyday life is a negligible fraction of what a physicist calls information. Entropy is defined using the *physical* concept of information.

Ref. 191 Finally, entropy is also *not* a measure for what in normal life is called the *complexity* of a situation. In fact, nobody has yet found a quantity describing this everyday notion.
Challenge 466 ny The task is surprisingly difficult. Have a try!

In summary, if you hear the term entropy used with a different meaning than $S = k \ln W$, beware. Somebody is trying to get you, probably with some ideology.

The minimum entropy of nature: the quantum of information

Before we complete our discussion of thermostatics we must point out in another way the importance of the Boltzmann constant k. We have seen that this constant appears whenever the granularity of matter plays a role; it expresses the fact that matter is made of small basic entities. The most striking way to put this statement is the following: *There is a smallest entropy in nature.* Indeed, for all systems, the entropy obeys

$$S \geqslant \frac{k}{2} . \qquad (96)$$

* When Max Planck went to Austria to search for the anonymous tomb of Boltzmann in order to get him buried in a proper grave, he inscribed the formula $S = k \ln W$ on the tombstone. (Which physicist would finance the tomb of another, nowadays?)

Challenge 465 ny * This is only approximate. Can you find the precise value?

Ref. 192
Ref. 193
This result is almost 100 years old; it was stated most clearly (with a different numerical factor) by the Hungarian–German physicist Leo Szilard. The same point was made by the French physicist Léon Brillouin (again with a different numerical factor). The statement can also be taken as the *definition* of the Boltzmann constant.

The existence of a smallest entropy in nature is a strong idea. It eliminates the possibility of the continuity of matter and also that of its fractality. A smallest entropy implies that matter is made of a finite number of small components. The limit to entropy expresses the fact that matter is made of particles.* The limit to entropy also shows that Galilean physics cannot be correct: Galilean physics assumes that arbitrarily small quantities do exist. The entropy limit is the first of several limits to motion that we will encounter until we finish the second part of our ascent. After we have found all limits, we can start the third and final part, leading to unification.

The existence of a smallest quantity implies a limit on the precision of measurement. Measurements cannot have infinite precision. This limitation is usually stated in the form of an indeterminacy relation. Indeed, the existence of a smallest entropy can be rephrased as an indeterminacy relation between the temperature T and the inner energy U of a system:

$$\Delta\frac{1}{T}\,\Delta U \geqslant \frac{k}{2}\,. \qquad (97)$$

Ref. 194
Page 996
Ref. 195
Ref. 193
This relation** was given by Niels Bohr; it was discussed by Werner Heisenberg, who called it one of the basic indeterminacy relations of nature. The Boltzmann constant (divided by 2) thus fixes the smallest possible entropy value in nature. For this reason, Gilles Cohen-Tannoudji calls it the *quantum of information* and Herbert Zimmermann calls it the *quantum of entropy*.

Page 656
The relation (97) points towards a more general pattern. For every minimum value for an observable, there is a corresponding indeterminacy relation. We will come across this several times in the rest of our adventure, most importantly in the case of the quantum of action and Heisenberg's indeterminacy relation.

The existence of a smallest entropy has numerous consequences. First of all, it sheds light on the third principle of thermodynamics. A smallest entropy implies that absolute zero is not achievable. Secondly, a smallest entropy explains why entropy values are finite instead of infinite. Thirdly, it fixes the absolute value of entropy for every system; in continuum physics, entropy, like energy, is only defined up to an additive constant. The entropy limit settles all these issues.

The existence of a minimum value for an observable implies that an indeterminacy relation appears for any two quantities whose product yields that observable. For example, entropy production rate and time are such a pair. Indeed, an indeterminacy relation connects the entropy production rate $P = \mathrm{d}S/\mathrm{d}t$ and the time t:

$$\Delta P\ \Delta t \geqslant \frac{k}{2}\,. \qquad (98)$$

* The minimum entropy implies that matter is made of tiny spheres; the minimum *action*, which we will encounter in quantum theory, implies that these spheres are actually small clouds.

Ref. 196
** It seems that the historical value for the right hand side, given by k, has to be corrected to $k/2$.

Form this and the previous relation (97) it is possible to deduce all of statistical physics, i.e., the precise theory of thermostatics and thermodynamics. We will not explore this further here. (Can you show that the zeroth principle follows from the existence of a smallest entropy?) We will limit ourselves to one of the cornerstones of thermodynamics: the second principle.

Ref. 195, Ref. 196

Challenge 467 ny

Why can't we remember the future?

> It's a poor sort of memory which only works backwards.
>
> Lewis Carroll (1832–1898), *Alice in Wonderland*

Page 42

When we first discussed time, we ignored the difference between past and future. But obviously, a difference exists, as we do not have the ability to remember the future. This is not a limitation of our brain alone. All the devices we have invented, such as tape recorders, photographic cameras, newspapers and books, only tell us about the past. Is there a way to build a video recorder with a 'future' button? Such a device would have to solve a deep problem: how would it distinguish between the near and the far future? It does not take much thought to see that any way to do this would conflict with the second principle of thermodynamics. That is unfortunate, as we would need precisely the same device to show that there is faster-than-light motion. Can you find the connection?

Challenge 468 ny

Challenge 469 ny

In summary, the future cannot be remembered because entropy in closed systems tends towards a maximum. Put even more simply, memory exists because the brain is made of many particles, and so the brain is limited to the past. However, for the most simple types of motion, when only a few particles are involved, the difference between past and future disappears. For few-particle systems, there is no difference between times gone by and times approaching. We could say that the future differs from the past only in our brain, or equivalently, only because of friction. Therefore the difference between the past and the future is not mentioned frequently in this walk, even though it is an essential part of our human experience. But the fun of the present adventure is precisely to overcome our limitations.

Is everything made of particles?

> A physicist is the atom's way of knowing about atoms.
>
> George Wald

Ref. 179

Historically, the study of statistical mechanics has been of fundamental importance for physics. It provided the first demonstration that physical objects are made of interacting particles. The story of this topic is in fact a long chain of arguments showing that all the properties we ascribe to objects, such as size, stiffness, colour, mass density, magnetism, thermal or electrical conductivity, result from the interaction of the many particles they consist of. The discovery that *all objects are made of interacting particles* has often been called the main result of modern science.

Page 213

How was this discovery made? Table 27 listed the main extensive quantities used in

Table 31 Some minimum flow values found in nature

Observation	Minimum value
Matter flow	one molecule, one atom or one particle
Volume flow	one molecule, one atom or one particle
Momentum flow	Planck's constant divided by wavelength
Angular momentum flow	Planck's constant
Chemical amount of substance	one molecule, one atom or one particle
Entropy flow	minimum entropy
Charge flow	elementary charge
Light flow	Planck's constant divided by wavelength

physics. Extensive quantities are able to flow. It turns out that all flows in nature are *composed* of elementary processes, as shown in Table 31. We have seen that the flow of mass, volume, charge, entropy and substance are composed. Later, quantum theory will show the same for the flow of linear and angular momentum. *All flows are made of particles.*

This success of this idea has led many people to generalize it to the statement: 'Everything we observe is made of parts.' This approach has been applied with success Ref. 180 to chemistry with molecules, materials science and geology with crystals, electricity with electrons, atoms with elementary particles, space with points, time with instants, light with photons, biology with cells, genetics with genes, neurology with neurons, mathematics with sets and relations, logic with elementary propositions, and even to linguistics with morphemes and phonemes. All these sciences have flourished on the idea that everything is made of related *parts*. The basic idea seems so self-evident that we find it Challenge 470 ny difficult even to formulate an alternative. Just try!

However, in the case of the *whole* of nature, the idea that nature is a sum of related Page 966 parts is incorrect. It turns out to be a prejudice, and a prejudice so entrenched that it retarded further developments in physics in the latter decades of the twentieth century. In particular, it does *not* apply to elementary particles or to space-time. Finding the correct description for the whole of nature is the biggest challenge of our adventure, as it requires a complete change in thinking habits. There is a lot of fun ahead.

> Jede Aussage über Komplexe läßt sich in eine Aussage über deren Bestandteile und in diejenigen Sätze zerlegen, welche die Komplexe vollständig beschreiben.*
>
> Ludwig Wittgenstein, *Tractatus*, 2.0201

Why stones can be neither smooth nor fractal, nor made of little hard balls

The exploration of temperature yields another interesting result. Researchers first studied gases, and measured how much energy was needed to heat them by 1 K. The result is

* Every statement about complexes can be resolved into a statement about their constituents and into the propositions that describe the complexes completely.

simple: all gases share only a few values, when the number of molecules N is taken into account. Monoatomic gases (in a container with constant volume) require $3Nk/2$, diatomic gases (and those with a linear molecule) $5Nk/2$, and almost all other gases $3Nk$, where $k = 1.4 \cdot 10^{-23}$ J/K is the Boltzmann constant. Page 221

The explanation of this result was soon forthcoming: each thermodynamic degree of freedom* contributes the energy $kT/2$ to the total energy, where T is the temperature. So the number of degrees of freedom in physical bodies is finite. Bodies are not continuous, nor are they fractals: if they were, their specific thermal energy would be infinite. Matter is indeed made of small basic entities.

All degrees of freedom contribute to the specific thermal energy. At least, this is what classical physics predicts. Solids, like stones, have 6 thermodynamic degrees of freedom and should show a specific thermal energy of $3Nk$. At high temperatures, this is indeed observed. But measurements of solids at room temperature yield lower values, and the lower the temperature, the lower the values become. Even gases show values lower than those just mentioned, when the temperature is sufficiently low. In other words, molecules and atoms behave differently at low energies: atoms are not immutable little hard balls. The deviation of these values is one of the first hints of quantum theory.

Curiosities and fun challenges about heat

Even though heat is disordered motion, it follows simple rules. Some of them are surprising.

▪ Compression of air increases its temperature. This is shown directly by the fire pump, a variation of a bicycle pump, shown in Figure 115. (For a working example, see the website http://www.tn.tudelft.nl/cdd). A match head at the bottom of an air pump made of transparent material is easily ignited by the compression of the air above it. The temperature of the air after compression is so high that the match head ignites spontaneously.

▪ If heat really is disordered motion of atoms, a big problem appears. When two atoms collide head-on, in the instant of smallest distance, neither atom has velocity. Where does the kinetic energy go? Obviously, it is transformed into potential energy. But that implies that atoms can be deformed, that they have internal structure, that they have parts, and thus that they can in principle be split. In short, if heat is disordered atomic motion, *atoms are not indivisible*! In the nineteenth century this argument was put forward in order to show that heat cannot be atomic motion, but must be some sort of fluid. But since we know that heat really is kinetic energy, atoms must indeed be divisible, even though their name means 'indivisible'. We do not need an expensive experiment to show this.

▪ Not only gases, but also most other materials expand when the temperature rises. As a result, the electrical wires supported by pylons hang much lower in summer than in winter. True? Challenge 471 n

Ref. 182 ▪ The following is a famous Fermi problem. Given that a human corpse cools down in four hours after death, what is the minimum number of calories needed per day in our food? Challenge 472 ny

* A *thermodynamic degree of freedom* is, for each particle in a system, the number of dimensions in which it can move plus the number of dimensions in which it is kept in a potential. Atoms in a solid have six, particles in monoatomic gases have only three; particles in diatomic gases or rigid linear molecules have five. The number of degrees of freedom of larger molecules depends on their shape. Ref. 181

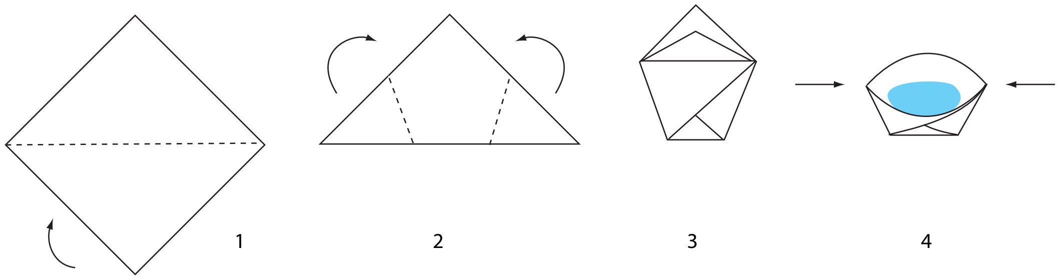


Figure 116 Can you to boil water in this paper cup?

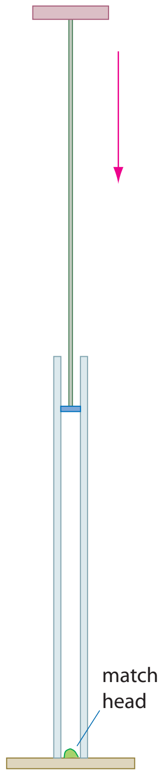


Figure 115 The fire pump

▪ The energy contained in thermal motion is not negligible. A 1 g bullet travelling at the speed of sound has a kinetic energy of only 0.01 kcal.

Challenge 473 n ▪ How does a typical, 1500 m^3 hot-air balloon work?

▪ Mixing 1 kg of water at 0°C and 1 kg of water at 100°C gives 2 kg of water at 50°C. What is the result of mixing 1 kg of *ice* at 0°C and 1 kg of water at 100°C? Challenge 474 ny

▪ The highest recorded air temperature in which a man has survived is 127°C. This was tested in 1775 in London, by the secretary of the Royal Society, Charles Blagden, together with a few friends, who remained in a room at that temperature for 45 minutes. Interestingly, the raw steak which he had taken in with him was cooked ('well done') when he and his friends left the room. What condition had to be strictly met in order to avoid cooking the people in the same way as the steak? Ref. 183 Challenge 475 n

Challenge 476 n ▪ Why does water boil at 99.975°C instead of 100°C?

Challenge 477 n ▪ Can you fill a bottle precisely with 1 ± 10^{-30} kg of water?

▪ If you do not like this text, here is a proposal. You can use the paper to make a cup, as shown in Figure 116, and boil water in it over an open flame. However, to succeed, you have to be a little careful. Can you find out in what way? Challenge 478 n

▪ One gram of fat, either butter or human fat, contains 38 kJ of chemical energy (or, in ancient units more familiar to nutritionists, 9 kcal). That is the same value as that of petrol. Why are people and butter less dangerous than petrol? Challenge 479 n

▪ In 1992, the Dutch physicist Martin van der Mark invented a loudspeaker which worked the heating of air by heating air with a laser beam. He demonstrated that with the right wavelength and with a suitable modulation of the intensity, a laser beam in air can generate sound, . The effect at the basis of this device, called the *photoacoustic effect*, appears in many materials. The best wavelength for air is in the infrared domain, on one of the few absorption lines of water vapour. In other words, a properly modulated infrared laser beam that shines through the air generates sound. The light can be emitted from a small matchbox-sized semiconductor laser hidden in the ceiling and shining downwards. The sound is emitted in all directions perpendicular to the beam. Since infrared laser light is not visible, Martin van der Mark thus invented an invisible loudspeaker! Unfortunately, the efficiency of present versions is still low, so that the power of the speaker is not yet sufficient for practical applications. Progress in laser technology should change this, so that in the future we should be able to hear sound that is emitted from the centre of an otherwise empty room.

▪ A famous exam question: How can you measure the height of a building with a barometer, a rope and a ruler? Find at least six different ways. Challenge 480 n

▪ What is the approximate probability that out of one million throws of a coin you get exactly 500 000 heads and as many tails? You may want to use Stirling's formula $n! \approx \sqrt{2\pi n}\,(n/e)^n$ to calculate the result.* Challenge 481 ny

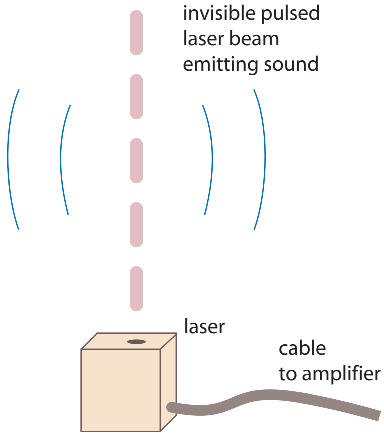


Figure 117 An invisible loudspeaker

▪ Does it make sense to talk about the entropy of the universe? Challenge 482 n

▪ Can a helium balloon lift the tank which filled it? Challenge 483 ny

▪ All friction processes, such as osmosis, diffusion, evaporation, or decay, are *slow*. They take a characteristic time. It turns out that any (macroscopic) process with a time-scale is irreversible. This is no real surprise: we know intuitively that undoing things always takes more time than doing them. That is again the second principle of thermodynamics.

▪ It turns out that *storing* information is possible with negligible entropy generation. However, *erasing* information requires entropy. This is the main reason why computers, as well as brains, require energy sources and cooling systems, even if their mechanisms would otherwise need no energy at all. Ref. 184

▪ When mixing hot rum and cold water, how does the increase in entropy due to the mixing compare with the entropy increase due to the temperature difference? Challenge 484 ny

▪ Why aren't there any small humans, e.g. 10 mm in size, as in many fairy tales? In fact, there are no warm-blooded animals of that size. Why not? Challenge 485 n

▪ Shining a light onto a body and repeatedly switching it on and off produces sound. This is called the *photoacoustic effect*, and is due to the thermal expansion of the material. By changing the frequency of the light, and measuring the intensity of the noise, one reveals a characteristic photoacoustic spectrum for the material. This method allows us to detect gas concentrations in air of one part in 10^9. It is used, among other methods, to study the gases emitted by plants. Plants emit methane, alcohol and acetaldehyde in small quantities; the photoacoustic effect can detect these gases and help us to understand the processes behind their emission.

▪ What is the rough probability that all oxygen molecules in the air would move away from a given city for a few minutes, killing all inhabitants? Challenge 486 ny

▪ If you pour a litre of water into the sea, stir thoroughly through all the oceans and then take out a litre of the mixture, how many of the original atoms will you find? Challenge 487 ny

▪ How long would you go on breathing in the room you are in if it were airtight? Challenge 488 ny

▪ What happens if you put some ash onto a piece of sugar and set fire to the whole? Challenge 489 ny

* There are many improvements to Stirling's formula. A simple one is $n! \approx \sqrt{(2n+1/3)\pi}\,(n/e)^n$. Another is $\sqrt{2\pi n}\,(n/e)^n e^{1/(12n+1)} < n! < \sqrt{2\pi n}\,(n/e)^n e^{1/(12n)}$.

(Warning: this is dangerous and not for kids.)

▪ Entropy calculations are often surprising. For a system of N particles with two states each, there are $W_{all} = 2^N$ states. For its most probable configuration, with exactly half the particles in one state, and the other half in the other state, we have $W_{max} = N!/((N/2)!)^2$. Now, for a macroscopic system of particles, we might typically have $N = 10^{24}$. That gives $W_{all} \gg W_{max}$; indeed, the former is 10^{12} times larger than the latter. On the other hand, we
Challenge 490 ny find that $\ln W_{all}$ and $\ln W_{max}$ agree for the first 20 digits! Even though the configuration with exactly half the particles in each state is much more rare than the general case, where
Challenge 491 ny the ratio is allowed to vary, the entropy turns out to be the same. Why?

▪ If heat is due to motion of atoms, our built-in senses of heat and cold are simply
Challenge 492 ny detectors of motion. How could they work?

By the way, the senses of smell and taste can also be seen as motion detectors, as they
Challenge 493 ny signal the presence of molecules flying around in air or in liquids. Do you agree?

▪ The Moon has an atmosphere, although an extremely thin one, consisting of sodium (Na) and potassium (K). This atmosphere has been detected up to nine Moon radii from its surface. The atmosphere of the Moon is generated at the surface by the ultraviolet
Challenge 494 n radiation from the Sun. Can you estimate the Moon's atmospheric density?

▪ Does it make sense to add a line in Table 27 for the quantity of physical action? A
Challenge 495 ny column? Why?

▪ Diffusion provides a length scale. For example, insects take in oxygen through their skin. As a result, the interiors of their bodies cannot be much more distant from the surface than about a centimetre. Can you list some other length scales in nature implied
Challenge 496 n by diffusion processes?

▪ Rising warm air is the reason why many insects are found in tall clouds in the evening. Many insects, especially that seek out blood in animals, are attracted to warm and humid air.

▪ Thermometers based on mercury can reach 750°C. How is this possible, given that
Challenge 497 n mercury boils at 357°C?

Challenge 498 n ▪ What does a burning candle look like in weightless conditions?

▪ It is possible to build a power station by building a large chimney, so that air heated by the Sun flows upwards in it, driving a turbine as it does so. It is also possible to make a power station by building a long vertical tube, and letting a gas such as ammonia rise into it which is then liquefied at the top by the low temperatures in the upper atmosphere; as it falls back down a second tube as a liquid – just like rain – it drives a turbine. Why are
Challenge 499 n such schemes, which are almost completely non-polluting, not used yet?

▪ One of the most surprising devices ever invented is the *Wirbelrohr* or Ranque–Hilsch vortex tube. By blowing compressed air at room temperature into it at its midpoint, two flows of air are formed at its ends. One is extremely cold, easily as low as −50°C, and one extremely hot, up to 200°C. No moving parts and no heating devices are found inside.
Challenge 500 n How does it work?

▪ It is easy to cook an egg in such a way that the white is hard but the yolk remains
Challenge 501 n liquid. Can you achieve the opposite?

▪ Thermoacoustic engines, pumps and refrigerators provide many strange and fascinating applications of heat. For example, it is possible to use loud sound in closed metal chambers to move heat from a cold place to a hot one. Such devices have few moving
Ref. 185 parts and are being studied in the hope of finding practical applications in the future.

▪ Does a closed few-particle system contradict the second principle of thermodynamics?
Challenge 502 ny

▪ What happens to entropy when gravitation is taken into account? We carefully left gravitation out of our discussion. In fact, many problems appear – just try to think about the issue. For example, Jakob Bekenstein has discovered that matter reaches its highest possible entropy when it forms a black hole. Can you confirm this?
Challenge 503 ny

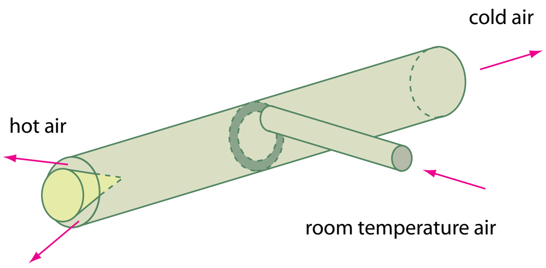


Figure 118 The Wirbelrohr or Ranque–Hilsch vortex tube

▪ The numerical values (but not the units!) of the Boltzmann constant $k = 1.38 \cdot 10^{-23}$ J/K and the combination h/ce agree in their exponent and in their first three digits, where h is Planck's constant and e is the electron charge. Can you dismiss this as mere coincidence?
Challenge 504 ny

Self-organization and chaos

> To speak of non-linear physics is like calling zoology the study of non-elephant animals.
>
> Stanislaw Ulam

Ref. 197

In our list of global descriptions of motion, the high point is the study of self-organization. Self-organization is the appearance of order. *Order* is a term that includes *shapes*, such as the complex symmetry of snowflakes; *patterns*, such as the stripes of zebras; and *cycles*, such as the creation of sound when singing. Every example of what we call *beauty* is a combination of shapes, patterns and cycles. (Do you agree?) Self-organization can thus be called the study of the origin of beauty.

Ref. 198
Challenge 505 n

The appearance of order is a general observation across nature. Fluids in particular exhibit many phenomena where order appears and disappears. Examples include the more or less regular flickering of a burning candle, the flapping of a flag in the wind, the regular stream of bubbles emerging from small irregularities in the surface of a champagne glass, and the regular or irregular dripping of a water tap.

Figures to be added

Figure 119 Examples of self-organization for sand

The appearance of order is found from the cell differentiation in an embryo inside a woman's body; the formation of colour patterns on tigers, tropical fish and butterflies; the symmetrical arrangements of flower petals; the formation of biological rhythms; and so on.

Table 32 Sand patterns in the sea and on land

Pattern	Period	Amplitude	Origin
sand banks	2 to 10 km	2 to 20 m	tides
sand waves	100 to 800 m	5 m	tides
megaribbles	1 m	0.1 m	tides
ribbles	5 cm	5 mm	waves
singing sand	95 to 105 Hz	up to 105 dB	wind on sand dunes, avalanches making the dune vibrate

All growth processes are self-organization phenomena. Have you ever pondered the incredible way in which teeth grow? A practically inorganic material forms shapes in the upper and the lower rows fitting exactly into each other. How this process is controlled is still a topic of research. Also the formation, before and after birth, of neural networks in the brain is another process of self-organization. Even the physical processes at the basis of thinking, involving changing electrical signals, is to be described in terms of self-organization.

Biological evolution is a special case of growth. Take the evolution of animal shapes. It turns out that snake tongues are forked because that is the most efficient shape for following chemical trails left by prey and other snakes of the same species. (Snakes smell with help of the tongue.) The fixed numbers of fingers in human hands or of petals of flowers are also consequences of self-organization.

Ref. 199
Page 650

Many problems of self-organization are mechanical problems: for example, the formation of mountain ranges when continents move, the creation of earthquakes, or the creation of regular cloud arrangements in the sky. It can be fascinating to ponder, during an otherwise boring flight, the mechanisms behind the formation of the clouds you see from the aeroplane.

Challenge 506 e

Studies into the conditions required for the appearance or disappearance of order have shown that their description requires only a few common concepts, independently of the details of the physical system. This is best seen looking at a few examples.

All the richness of self-organization reveals itself in the study of plain sand. Why do sand dunes have ripples, as does the sand floor at the bottom of the sea? We can also study how avalanches occur on steep heaps of sand and how sand behaves in hourglasses, in mixers, or in vibrating containers. The results are often surprising. For example, as recently as 1996 Paul Umbanhowar and his colleagues found that when a flat container holding tiny bronze balls (around 0.165 mm in diameter) is shaken up and down in vacuum at certain frequencies, the surface of this bronze 'sand' forms stable heaps. They are shown in Figure 120. These heaps, so-called *oscillons*, also bob up and down. Oscillons can move and interact with one another.

Ref. 200

Oscillons in sand are simple example for a general effect in nature: *discrete* systems with nonlinear interactions can exhibit localized excitations. This fascinating topic is just beginning to be researched. It might well be that it will yield results relevant to our understanding of elementary particles.

Ref. 201

Sand shows many other pattern-forming processes. A mixture of sand and sugar, when poured onto a heap, forms regular layered structures that in cross section look like zebra stripes. Horizontally rotating cylinders with binary mixtures inside them separate the mixture out over time. Or take a container with two compartments separated by a 1 cm wall. Fill both halves with sand and rapidly shake the whole container with a machine. Over time, all the sand will spontaneously accumulate in one half of the container. As another example of self-organization in sand, people have studied the various types of sand dunes that 'sing' when the wind blows over them. In fact, the behaviour of sand and dust is proving to be such a beautiful and fascinating topic that the prospect of each human returning dust does not look so grim after all.

Ref. 202

Figure 120 Oscillons formed by shaken bronze balls; horizontal size is about 2 cm (© Paul Umbanhowar)

Another simple and beautiful example of self-organization is the effect discovered in 1999 by Karsten Kötter and his group. They found that the behaviour of a set of spheres swirled in a dish depends on the number of spheres used. Usually, all the spheres get continuously mixed up. But for certain 'magic' numbers, such as 21, stable ring patterns emerge, for which the outside spheres remain outside and the inside ones remain inside. The rings, best seen by colouring the spheres, are shown in Figure 121.

Ref. 203

These and many other studies of self-organizing systems have changed our understanding of nature in a number of ways. First of all, they have shown that patterns and shapes are similar to cycles: all are due to motion. Without motion, and thus without history, there is no order, neither patterns nor shapes. Every pattern has a history; every pattern is a result of motion.

Ref. 204

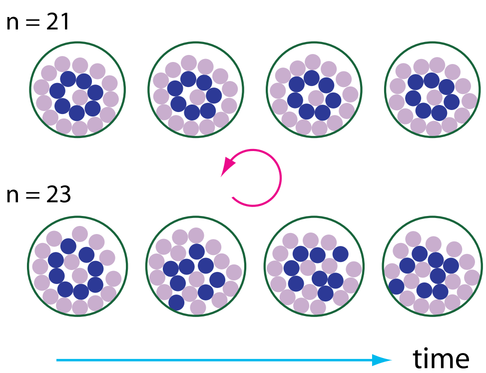


Figure 121 Magic numbers: 21 spheres, when swirled in a dish, behave differently from non-magic numbers, like 23, of spheres (redrawn from photographs, © Karsten Kötter)

Secondly, patterns, shapes and cycles are due to the organized motion of large numbers of small constituents. Systems which self-organize are always composite: they are *cooperative structures*.

Thirdly, all these systems obey evolution equations which are *nonlinear* in the configuration variables. Linear systems do not self-organize. Many self-organizing systems also show *chaotic* motion.

Fourthly, the appearance and disappearance of order depends on the strength of a driving force, the so-called *order parameter*. Often, chaotic motion appears when the driving is increased beyond the value necessary for the appearance of order. An example of chaotic motion is turbulence, which appears when the order parameter, which is proportional to the speed of the fluid, is increased to high values.

Moreover, all order and all structure appears when two general types of motion compete with each other, namely a 'driving', energy-adding process, and a 'dissipating', braking mechanism. Thermodynamics plays a role in all self-organization. Self-organizing systems are always *dissipative systems*, and are always far from equilibrium. When the driving and the dissipation are of the same order of magnitude, and when the key behaviour of the system is not a linear function of the driving action, order may appear.*

All self-organizing systems at the onset of order appearance can be described by equations for the pattern amplitude A of the general form

$$\frac{\partial A(t,x)}{\partial t} = \lambda A - \mu |A|^2 A + \kappa\, \Delta A + \text{higher orders}\,. \qquad (99)$$

Here, the – possibly complex – observable A is the one that appears when order appears, such as the oscillation amplitude or the pattern amplitude. The first term λA is the driving term, in which λ is a parameter describing the strength of the driving. The next term is a typical nonlinearity in A, with μ a parameter that describes its strength, and the third term $\kappa\, \Delta A = \kappa(\partial^2 A/\partial x^2 + \partial^2 A/\partial y^2 + \partial^2 A/\partial z^2)$ is a typical dissipative (and diffusive) term.

One can distinguish two main situations. In cases where the dissipative term plays no role ($\kappa = 0$), one finds that when the driving parameter λ increases above zero, a *temporal* oscillation appears, i.e. a stable cycle with non-vanishing amplitude. In cases where the diffusive term does play a role, equation (99) describes how an amplitude for a *spatial* oscillation appears when the driving parameter λ becomes positive, as the solution $A = 0$ then becomes spatially unstable.

Challenge 507 ny

Challenge 508 ny

In both cases, the onset of order is called a *bifurcation*, because at this critical value of the driving parameter λ the situation with amplitude zero, i.e. the homogeneous (or unordered) state, becomes unstable, and the ordered state becomes stable. *In nonlinear systems, order is stable.* This is the main conceptual result of the field. Equation (99) and its numerous variations allow us to describe many phenomena, ranging from spirals, waves, hexagonal patterns, and topological defects, to some forms of turbulence. For every physical system under study, the main task is to distil the observable A and the parameters λ, μ and κ from the underlying physical processes.

Ref. 205

Self-organization is a vast field which is yielding new results almost by the week. To discover new topics of study, it is often sufficient to keep one's eye open; most effects are comprehensible without advanced mathematics. Good hunting!

Challenge 509 ny

Most systems that show self-organization also show another type of motion. When the driving parameter of a self-organizing system is increased to higher and higher values,

* To describe the 'mystery' of human life, terms like 'fire', 'river' or 'tree' are often used as analogies. These are all examples of self-organized systems: they have many degrees of freedom, have competing driving and braking forces, depend critically on their initial conditions, show chaos and irregular behaviour, and sometimes show cycles and regular behaviour. Humans and human life resemble them in all these respects; thus there is a solid basis to their use as metaphors. We could even go further and speculate that pure beauty *is* pure self-organization. The lack of beauty indeed often results from a disturbed equilibrium between external braking and external driving.

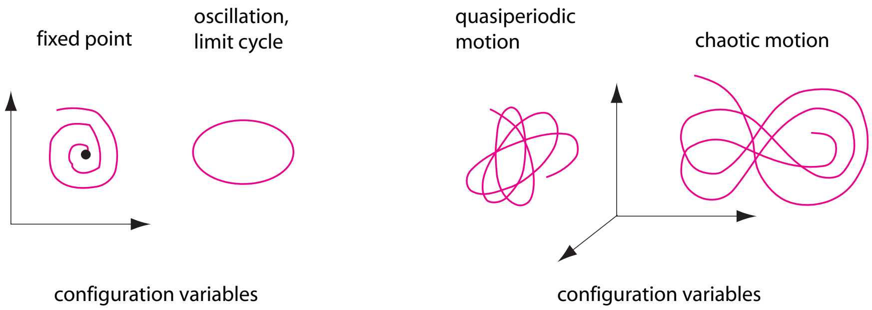


Figure 122 Examples of different types of motion in configuration space

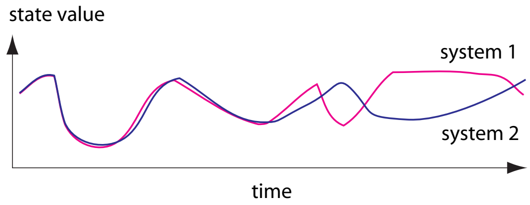


Figure 123 Sensitivity to initial conditions

order becomes more and more irregular, and in the end one usually finds chaos. For physicists, chaotic motion is the most irregular type of motion.* Chaos can be defined independently of self-organization, namely as that motion of systems for which small changes in initial conditions evolve into large changes of the motion (exponentially with time), as shown in Figure 123. More precisely, *chaos* is irregular motion characterized by a positive *Lyapounov exponent*. The weather is such a system, as are dripping water-taps, the fall of dice, and many other common systems. For example, research on the mechanisms by which the heart beat is generated has shown that the heart is not an oscillator, but a chaotic system with irregular cycles. This allows the heart to be continuously ready for demands for changes in beat rate which arise once the body needs to increase or decrease its efforts. Ref. 140

Incidentally, can you give a simple argument to show that the so-called *butterfly effect* does not exist? Challenge 510 n This 'effect' is often cited in newspapers: the claim is that nonlinearities imply that a small change in initial conditions can lead to large effects; thus a butterfly wing beat is alleged to be able to induce a tornado. Even though nonlinearities do indeed lead to growth of disturbances, the butterfly effect has never been observed; it does not exist.

There is chaotic motion also in machines: chaos appears in the motion of trains on the rails, in gear mechanisms, and in fire-fighter's hoses. The precise study of the motion in a zippo cigarette lighter will probably also yield an example of chaos. Challenge 511 ny The mathematical

* On the topic of chaos, see the beautiful book by H.-O. Peitgen, H. Jürgens & D. Saupe, *Chaos and Fractals*, Springer Verlag, 1992. It includes stunning pictures, the necessary mathematical background, and some computer programs allowing personal exploration of the topic. 'Chaos' is an old word: according to Greek mythology, the first goddess, Gaia, i.e. the Earth, emerged from the chaos existing at the beginning.

description of chaos – simple for some textbook examples, but extremely involved for others – remains an important topic of research.

All the steps from disorder to order, quasiperiodicity and finally to chaos, are examples of self-organization. These types of motion, illustrated in Figure 122, are observed in many fluid systems. Their study should lead, one day, to a deeper understanding of the mysteries Ref. 206 of turbulence. Despite the fascination of this topic, we will not explore it further, because it does not lead towards the top of Motion Mountain.

But self-organization is of interest also for a more general reason. It is sometimes said that our ability to formulate the patterns or rules of nature from observation does not imply the ability to predict *all* observations from these rules. According to this view, so-called 'emergent' properties exist, i.e. properties appearing in complex systems as something *new* that cannot be deduced from the properties of their parts and their interactions. (The ideological backdrop to this view is obvious; it is the last attempt to fight the determinism.) The study of self-organization has definitely settled this debate. The properties of water molecules do allow us to predict Niagara Falls.* Similarly, the diffusion of signal molecules do determine the development of a single cell into a full human being: in particular, cooperative phenomena determine the places where arms and legs are formed; they ensure the (approximate) right–left symmetry of human bodies, prevent mix-ups of connections when the cells in the retina are wired to the brain, and explain the fur patterns on zebras and leopards, to cite only a few examples. Similarly, the mechanisms at the origin of the heart beat and many other cycles have been deciphered.

Self-organization provides general principles which allow us in principle to predict the behaviour of complex systems of any kind. They are presently being applied to the most complex system in the known universe: the human brain. The details of how it learns to coordinate the motion of the body, and how it extracts information from the images in the eye, are being studied intensely. The ongoing work in this domain is fascinating. If Challenge 513 ny you plan to become a scientist, consider taking this path.

Such studies provide the final arguments that confirm what J. Offrey de la Mettrie in 1748 stated and explored in his famous book *L'homme machine*: humans are complex machines. Indeed, the lack of understanding of complex systems in the past was due mainly to the restrictive teaching of the subject of motion, which usually concentrated – as we do in this walk – on examples of motion in *simple* systems. Even though the subject of self-organization provides fascinating insights, and will do so for many years to come, we now leave it. We continue with our own adventure exploring the basics of motion.**

Curiosities and fun challenges about self-organization

Every example of a pattern or of beauty contains a physical challenge:

- All icicles have a wavy surface, with a crest-to-crest distance of about 1 cm, as shown in Figure 124. The distance is determined by the interplay between water flow and surface

She then gave birth to the other gods, the animals and the first humans.

* Already small versions of Niagara Falls, namely dripping water taps, show a large range of cooperative Ref. 207 phenomena, including the chaotic, i.e. non-periodic, fall of water drops. This happens when the water flow Challenge 512 ny has the correct value, as you can verify in your own kitchen. Several cooperative fluid phenomena have been simulated even on the molecular level.

Ref. 208 ** An important case of self-organization is *humour*.

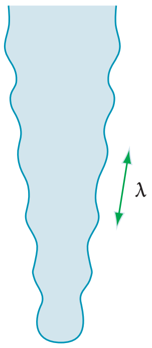


Figure 124 The wavy surface of icicles

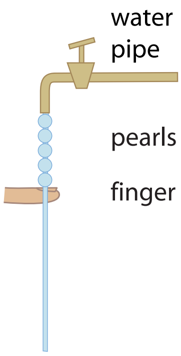


Figure 125 Water pearls

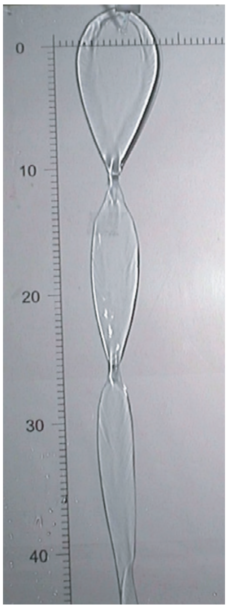


Figure 126 A braiding water stream (© Vakhtang Putkaradze)

Challenge 514 ny cooling. How?

▪ When wine is made to swirl in a wine glass, after the motion has calmed down, the wine flowing down the glass walls forms little arcs. Can you explain in a few words what forms them? Challenge 515 ny

▪ How does the average distance between cars parked along a street change over time, assuming a constant rate of cars leaving and arriving? Challenge 516 d

▪ When a fine stream of water leaves a water tap, putting a finger in the stream leads to a wavy shape, as shown in Figure 125. Why? Challenge 517 d

▪ When water emerges from a oblong opening, the stream forms a braid pattern, as shown in Figure 126. This effect results from the interplay and competition between inertia and surface tension: inertia tends to widen the stream, while surface tension tends to narrow it. Predicting the distance from one narrow region to the next is still a topic of research. Ref. 186

If the experiment is done in free air, without a plate, one usually observes an additional effect: there is a *chiral* braiding at the narrow regions, induced by the asymmetries of the water flow. You can observe this effect in the toilet! Scientific curiosity knows no limits: are you a right-turner or a left-turner, or both? On every day? Challenge 518 ny

▪ Gerhard Müller has discovered a simple but beautiful way to observe self-organization in solids. His system also provides a model for a famous geological process, the formation of hexagonal columns in basalt, such as the Devil's Staircase in Ireland. Similar formations are found in many other places of the Earth. Just take some rice flour

Ref. 187 Challenge 519 e or corn starch, mix it with about half the same amount of water, put the mixture into a pan and dry it with a lamp. Hexagonal columns form. The analogy works because the drying of starch and the cooling of lava are diffusive processes governed by the same equations, because the boundary conditions are the same, and because both materials respond with a small reduction in volume.

▪ Water flow in pipes can be laminar (smooth) or turbulent (irregular and disordered). The transition depends on the diameter d of the pipe and the speed v of the water. The transition usually happens when the so-called *Reynolds number* – defined as $R = vd/\eta$ (η being the kinematic viscosity of the water, around $1\,\mathrm{mm^2/s}$) – becomes greater than about 2000. However, careful experiments show that with proper handling, laminar flows can be produced up to $R = 100\,000$. A linear analysis of the equations of motion of the fluid, the Navier–Stokes equations, even predicts stability of laminar flow for *all* Reynolds numbers. This riddle was solved only in the years 2003 and 2004. First, a complex mathematical analysis showed that the laminar flow is not always stable, and that the transition to turbulence in a long pipe occurs with travelling waves. Then, in 2004, careful experiments showed that these travelling waves indeed appear when water is flowing through a pipe at large Reynolds numbers. Ref. 209

5. From the limitations of physics to the limits of motion

> I only know that I know nothing.
> Socrates (470–399 BCE), as cited by Plato

Socrates' saying applies also to Galilean physics, despite its general success in engineering and in the description of everyday life. We will now give a short overview of the limitations of the field.

Research topics in classical dynamics

Even though the science of mechanics is now several hundred years old, research into its details is still continuing.

▪ We have already mentioned above the issue of the stability of the solar system. The long-term future of the planets is unknown. In general, the behaviour of few-body systems interacting through gravitation is still a research topic of mathematical physics. Answering the simple question of how long a given set of bodies gravitating around each other will stay together is a formidable challenge. The history of this so-called *many-body problem* is long and involved. Interesting progress has been achieved, but the final answer still eludes us. Ref. 210

▪ Many challenges remain in the fields of self-organization, of nonlinear evolution equations, and of chaotic motion; and they motivate numerous researchers in mathematics, physics, chemistry, biology, medicine and the other sciences.

▪ Perhaps the toughest of all problems in physics is how to describe *turbulence*. When the young Werner Heisenberg was asked to continue research on turbulence, he refused – rightly so – saying it was too difficult; he turned to something easier and discovered quantum mechanics instead. Turbulence is such a vast topic, with many of its concepts still not settled, that despite the number and importance of its applications, only now,

Ref. 211 at the beginning of the twenty-first century, are its secrets beginning to be unravelled. It is thought that the equations of motion describing fluids, the so-called *Navier–Stokes equations*, are sufficient to understand turbulence.* But the mathematics behind them is mind-boggling. There is even a prize of one million dollars offered by the Clay Mathematics Institute for the completion of certain steps on the way to solving the equations.

What is contact?

> Democritus declared that there is a unique sort of motion: that ensuing from collision.
> Simplicius, *Commentary on the Physics of Aristotle*, 42, 10

Ref. 212

Of the questions unanswered by classical physics, the details of contact and collisions are among the most pressing. Indeed, we defined mass in terms of velocity changes during Page 72 collisions. But why do objects change their motion in such instances? Why are collisions between two balls made of chewing gum different from those between two stainless-steel balls? What happens during those moments of contact?

Contact is related to material properties, which in turn influence motion in a complex way. The complexity is such that the sciences of material properties developed independently from the rest of physics for a long time; for example, the techniques of metallurgy (often called the oldest science of all) of chemistry and of cooking were related to the properties of motion only in the twentieth century, after having been independently pursued for thousands of years. Since material properties determine the essence of contact, we *need* knowledge about matter and about materials to understand the notion of mass, and thus of motion. The second part of our mountain ascent will reveal these connections.

Precision and accuracy

When we started climbing Motion Mountain, we stated that to gain height means to increase the *precision* of our description of nature. To make even this statement itself more precise, we distinguish between two terms: *precision* is the degree of reproducibility; *accuracy* is the degree of correspondence to the actual situation. Both concepts apply to measurements,* to statements and to physical concepts.

Appendix B Atpresent, the record number of digits ever measured for a physical quantity is 14. Why so few? Classical physics doesn't provide an answer. What is the maximum number of digits we can expect in measurements; what determines it; and how can we achieve it? These questions are still open at this point in our ascent; they will be covered in the second part of it.

On the other hand, statements with false accuracy abound. What should we think of a car company – Ford – who claim that the drag coefficient c_w of a certain model Challenge 520 n is 0.375? Or of the official claim that the world record in fuel consumption for cars is

* They are named after Claude Navier (b. 1785 Dijon, d. 1836 Paris), important French engineer and bridge builder, and Georges Gabriel Stokes (b. 1819 Skreen, d. 1903 Cambridge), important Irish physicist and mathematician.

* For measurements, both precision and accuracy are best described by their *standard deviation*, as explained in Appendix B, on page 1069.

2315.473 km/l? Or of the statement that 70.3 % of all citizens share a certain opinion? One lesson we learn from investigations into measurement errors is that we should never provide more digits for a result than we can put our hand into fire for.

Challenge 521 n

Is it possible to draw or produce a rectangle for which the ratio of lengths is a real number, e.g. of the form 0.131520091514001315211420010914..., whose digits encode a book? (A simple method would code a space as 00, the letter 'a' as 01, 'b' as 02, 'c' as 03, etc. Even more interestingly, could the number be printed inside its own book?)

In our walk we aim for precision and accuracy, while avoiding false accuracy. Therefore, concepts have mainly to be *precise*, and descriptions have to be *accurate*. Any inaccuracy is a proof of lack of understanding. To put it bluntly, 'inaccurate' means *wrong*. Increasing the accuracy and precision of our description of nature implies leaving behind us all the mistakes we have made so far. That is our aim in the following.

Can all of nature be described in a book?

Let us have some fun with a paradox related to our adventure. If a perfect physics publication describing all of nature existed, it must also describe itself, its own production – including its author – and most important of all, its own contents. Is this possible? Using the concept of information, we can state that such a book should contain all information contained in the universe, including the information in the book itself. Is this possible?

If nature requires an *infinitely* long book to be fully described, such a publication obviously cannot exist. In this case, only approximate descriptions of nature are possible.

If nature requires a *finite* amount of information for its description, then the universe cannot contain more information than is already contained in the book. This would imply that the rest of the universe would not add to the information already contained in the book. It seems that the entropy of the book and the entropy of the universe must be similar. This is possible, but seems somewhat unlikely.

We note that the answer to this puzzle also implies the answer to another puzzle: whether a brain can contain a full description of nature. In other words, the question is: can humans understand nature? We do believe so. In other words, we seem to believe something rather unlikely: that the universe does not contain more information than what our brain could contain or even contains already. However, this conclusion is not correct. The terms 'universe' and 'information' are not used correctly in this reasoning, as you might want to verify. We will solve this puzzle later in our adventure. Until then, do make up your own mind.

Page 969
Challenge 522 e

Why is measurement possible?

In the description of gravity given so far, the one that everybody learns – or should learn – at school, acceleration is connected to mass and distance via $a = GM/r^2$. That's all. But this simplicity is deceiving. In order to check whether this description is correct, we have to measure lengths and times. However, it is *impossible* to measure lengths and time intervals with any clock or any ruler based on the gravitational interaction alone! Try to conceive such an apparatus and you will be inevitably be disappointed. You always need a non-gravitational method to start and stop the stopwatch. Similarly, when you measure length, e.g. of a table, you have to hold a ruler or some other device near it. The interaction

Challenge 523 n

necessary to line up the ruler and the table cannot be gravitational.

A similar limitation applies even to mass measurements. Try to measure mass using gravitation alone. Any scale or balance needs other – usually mechanical, electromagnetic or optical – interactions to achieve its function. Can you confirm that the same applies to speed and to angle measurements? In summary, whatever method we use, *in order to measure velocity, length, time, and mass, interactions other than gravity are needed.* Our ability to measure shows that gravity is not all there is.

Challenge 524 n
Challenge 525 n

Is motion unlimited?

Galilean physics does not explain the ability to measure. In fact, it does not even explain the existence of standards. Why do objects have fixed lengths? Why do clocks work with regularity? Galilean physics cannot explain these observations.

Galilean physics also makes no clear statements on the universe as a whole. It seems to suggest that it is infinite. Finitude does not fit with the Galilean description of motion. Galilean physics is thus limited in its explanations because it disregards the limits of motion.

We also note that the existence of infinite speeds in nature would not allow us to define time sequences. Clocks would then be impossible. In other words, a description of nature that allows unlimited speeds is not precise. Precision requires limits. To achieve the highest possible precision, we need to discover all limits to motion. So far, we have discovered only one: there is a smallest entropy. We now turn to another, more striking one: the limit for speed. To understand this limit, we will explore the most rapid motion we know: the motion of light.

Bibliography

127 The beautiful story of the south-pointing carriage is told in Appendix B of James Foster & J.D. Nightingale, *A Short Course in General Relativity*, Springer Verlag, 2nd edition, 1998. Such carriages have existed in China, as told by the great sinologist Joseph Needham, but their construction is unknown. The carriage described by Foster and Nightingale is the one reconstructed in 1947 by George Lancaster, a British engineer. Cited on page 159.

128 See for example Z. Ghahramani, Building blocks of movement, *Nature* **407**, pp. 682–683, 2000. Researchers in robot control are also interested in such topics. Cited on page 159.

129 G. Gutierrez, C. Fehr, A. Calzadilla & D. Figueroa, Fluid flow up the wall of a spinning egg, *American Journal of Physics* **66**, pp. 442–445, 1998. Cited on page 159.

130 A historical account is given in Wolfgang Yourgray & Stanley Mandelstam, *Variational Principles in Dynamics and Quantum Theory*, Dover, 1968. Cited on pages 161 and 167.

131 Max Päsler, *Prinzipe der Mechanik*, Walter de Gruyter & Co., 1968. Cited on page 165.

132 The relations between possible Lagrangians are explained by Herbert Goldstein, *Classical Mechanics*, 2nd edition, Addison-Wesley, 1980. Cited on page 166.

133 C.G. Gray, G. Karl & V.A. Novikov, From Maupertius to Schrödinger. Quantization of classical variational principles, *American Journal of Physics* **67**, pp. 959–961, 1999. Cited on page 167.

134 G.B. West, J.H. Brown & B.J. Enquist, A general model for the origin of allometric scaling laws in biology, *Science* **276**, pp. 122–126, 4 April 1997, with a comment on page 34 of the same issue. The rules governing branching properties of blood vessels, of lymph systems and of vessel systems in plants are explained. For more about plants, see also the paper G.B. West, J.H. Brown & B.J. Enquist, A general model for the structure and allometry of plant vascular systems, *Nature* **400**, pp. 664–667, 1999. Cited on page 172.

135 J.R. Banavar, A. Martin & A. Rinaldo, Size and form in efficient transportation networks, *Nature* **399**, pp. 130–132, 1999. Cited on page 172.

136 The Hemingway statement is quoted by Marlene Dietrich in Aaron E. Hotchner, *Papa Hemingway*, Random House, 1966, in part 1, chapter 1. Cited on page 167.

137 J.A. Moore, An innovation in physics instruction for nonscience majors, *American Journal of Physics* **46**, pp. 607–612, 1978. Cited on page 168.

138 See e.g. Alan P. Boss, Extrasolar planets, *Physics Today* **49**, pp. 32–38. September 1996. The most recent information can be found at the 'Extrasolar Planet Encyclopaedia' maintained at http://www.obspm.fr/planets by Jean Schneider at the Observatoire de Paris. Cited on page 170.

139 A good review article is by David W. Hughes, Comets and Asteroids, *Contemporary Physics* **35**, pp. 75–93, 1994. Cited on page 170.

140 An idea of the fascinating mechanisms at the basis of the heart beat is given by A. Babloyantz & A. Destexhe, Is the normal heart a periodic oscillator?, *Biological Cybernetics* **58**, pp. 203–211, 1989. Cited on page 236.

141 See e.g. B. Bower, A child's theory of mind, *Science News* **144**, pp. 40–41. Cited on page 175.

142 The most beautiful book on this topic is the text by Branko Grünbaum & G.C. Shephard, *Tilings and Patterns*, W.H. Freeman and Company, New York, 1987. It has been translated into several languages and republished several times. Cited on page 176.

143 U. Niederer, The maximal kinematical invariance group of the free Schrödinger equation, *Helvetica Physica Acta* **45**, pp. 802–810, 1972. See also the introduction by O. Jahn & V.V. Sreedhar, The maximal invariance group of Newton's equations for a free point particle, http://www.arxiv.org/abs/math-ph/0102011. Cited on page 182.

144 The story is told in the interesting biography of Einstein by A. Pais, *'Subtle is the Lord...' – The Science and the Life of Albert Einstein*, Oxford University Press, 1982. Cited on page 183.

145 W. Zürn & R. Widmer-Schnidrig, Globale Eigenschwingungen der Erde, *Physik Journal* **1**, pp. 49–55, 2002. Cited on page 188.

146 N. Gauthier, What happens to energy and momentum when two oppositely-moving wave pulses overlap?, *American Journal of Physics* **71**, pp. 787–790, 2003. Cited on page 191.

147 An informative and modern summary of present research about the ear and the details of its function is http://www.physicsweb.org/article/world/15/5/8. Cited on page 193.

148 A.L. Hodgkin & A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *Journal of Physiology* **117**, pp. 500–544, 1952. This famous paper of theoretical biology earned the authors the Nobel Prize in Medicine in 1963. Cited on page 195.

149 For more about the ocean sound channel, see the novel by Tom Clancy, *The Hunt for Red October*. See also the physics script by R.A. Muller, Government secrets of the oceans, atmosphere, and UFOs, http://web.archive.org/web/*/http://muller.lbl.gov/teaching/Physics10/chapters/9-SecretsofUFOs.html 2001. Cited on page 199.

150 B. Wilson, R.S. Batty & L.M. Dill, Pacific and Atlantic herring produce burst pulse sounds, *Biology Letters* **271**, number S3, 7 February 2004. Cited on page 199.

151 N.J. Zabusky & M.D. Kruskal, Interaction of solitons in a collisionless plasma and the recurrence of initial states, *Physical Review Letters* **15**, pp. 240–243, 1965. Cited on page 197.

152 O. Muskens, De kortste knal ter wereld, *nederlands tijdschrift voor natuurkunde* pp. 70–73, 2005. Cited on page 197.

153 T. Filippov, *The Versatile Soliton*, Springer Verlag, 2000. See also J.S. Russel, Report of the Fourteenth Meeting of the British Association for the Advancement of Science, Murray, London, 1844, pp. 311–390. Cited on pages 195 and 197.

154 See for example the article by G. Fritsch, Infraschall, *Physik in unserer Zeit* **13**, pp. 104–110, 1982. Cited on page 200.

155 Wavelet transformations were developed by the French mathematicians Alex Grossmann, Jean Morlet and Thierry Paul. The basic paper is A. Grossmann, J. Morlet & T. Paul, Integral transforms associated to square integrable representations, *Journal of Mathematical Physics* **26**, pp. 2473–2479, 1985. For a modern introduction, see Stéphane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999. Cited on page 200.

156 Jay Ingram, *The Velocity of Honey*, Viking, 2003. Cited on page 200.

157 M. Ausloos, *Proceedings of the Royal Society in London A* **400**, pp. 331–350, 1985. Cited on page 201.

158 T.A. McMahob & J. Tyler Bonner, *Form und Leben – Konstruktion vom Reißbrett der Natur*, Spektrum Verlag, 1985. Cited on page 203.

159 G.W. Koch, S.C. Sillett, G.M. Jennings & S.D. Davis, The limits to tree height, *Nature* **428**, pp. 851–854, 2004. Cited on page 204.

160 A simple article explaining the tallness of trees is A. Mineyev, Trees worthy of Paul Bunyan, *Quantum* pp. 4–10, January–February 1994. (Paul Bunyan is a mythical giant lumberjack who is the hero of the early frontier pioneers in the United States.) Note that the transport of liquids in trees sets no limits on their height, since water is pumped up along tree stems (except in spring, when it is pumped up from the roots) by evaporation from the leaves. This works almost without limits because water columns, when nucleation is carefully avoided, can be put under tensile stresses of over 100 bar, corresponding to 1000 m. See also P. Nobel, *Plant Physiology*, Academic Press, 2nd Edition, 1999. Cited on page 204.

161 Such information can be taken from the excellent overview article by M.F. Ashby, On the engineering properties of materials, *Acta Metallurgica* **37**, pp. 1273–1293, 1989. The article explains the various general criteria which determine the selection of materials, and gives numerous tables to guide the selection. Cited on page 204.

162 For a photograph of a single Barium *atom* – named Astrid – see Hans Dehmelt, Experiments with an isolated subatomic particle at rest, *Reviews of Modern Physics* **62**, pp. 525–530, 1990. For an earlier photograph of a Barium *ion*, see W. Neuhauser, M. Hohenstatt, P.E. Toschek & H. Dehmelt, Localized visible Ba^+ mono-ion oscillator, *Physical Review A* **22**, pp. 1137–1140, 1980. Cited on page 206.

163 Holograms of atoms were first produced by Hans-Werner Fink & al., Atomic resolution

in lens-less low-energy electron holography, *Physical Review Letters* **67**, pp. 1543–1546, 1991. Cited on page 206.

164 A single–atom laser was built in 1994 by K. An, J.J. Childs, R.R. Dasari & M.S. Feld, Microlaser: a laser with one atom in an optical resonator, *Physical Review Letters* **73**, p. 3375, 1994. Cited on page 206.

165 See for example C. Schiller, A.A. Koomans, T.L. van Rooy, C. Schönenberger & H.B. Elswijk, Decapitation of tungsten field emitter tips during sputter sharpening, *Surface Science Letters* **339**, pp. L925–L930, 1996. Cited on page 207.

166 U. Weierstall & J.C.H. Spence, An STM with time-of-flight analyzer for atomic species identification, MSA 2000, Philadelphia, *Microscopy and Microanalysis* **6**, Supplement 2, p. 718, 2000. Cited on page 207.

167 Lydéric Bocquet, The physics of stone skipping, *American Journal of Physics* **17**, pp. 150–155, 2003. The present recod holder is Kurt Steiner, with 40 skips. See http://web.archive.org/web/20041010013318/http://www.riverromp.com/skipping/steiner.htm and http://www.stoneskipping.com. The site http://www.yeeha.net/nassa/guin/g2.html is by the a previous world record holder, Jerdome Coleman–McGhee. Cited on page 208.

168 S.F. Kistler & L.E. Scriven, The teapot effect: sheetforming flows with deflection, wetting, and hysteresis, *Journal of Fluid Mechanics* **263**, pp. 19–62, 1994. Cited on page 210.

169 P. Krehl, S. Engemann & D. Schwenkel, The puzzle of whip cracking – uncovered by a correlation of whip-tip kinematics with shock wave emission, *Shock Waves* **8**, pp. 1–9, 1998. The authors used high-speed cameras to study the motion of the whip. A new aspect has been added by A. Goriely & T. McMillen, Shape of a cracking whip, *Physical Review Letters* **88**, p. 244301, 2002. This article focuses on the tapered shape of the whip. However, the neglection of the tuft – a piece at the end of the whip which is required to make it crack – in the latter paper shows that there is more to be discovered still. Cited on page 211.

170 S. Dorbolo, H. Caps & N. Vandewalle, Fluid instabilities in the birth and death of antibubbles, *New Journal of Physics* **5**, p. 161, 2003. Cited on page 211.

171 Thermostatics is difficult to learn also because it was not discovered in a systematic way. See C. Truesdell, *The Tragicomical History of Thermodynamics 1822–1854*, Springer Verlag, 1980. An excellent advanced textbook on thermostatics and thermodynamics is Linda Reichl, *A Modern Course in Statistical Physics*, Wiley, 2nd edition, 1998. Cited on page 214.

172 The famous paper on Brownian motion which contributed so much to Einstein's fame is A. Einstein, Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, *Annalen der Physik* **17**, pp. 549–560, 1905. In the following years, Einstein wrote a series of further papers elaborating on this topic. For example, he published his 1905 Ph.D. thesis as A. Einstein, Eine neue Bestimmung der Moleküldimensionen, *Annalen der Physik* **19**, pp. 289–306, 1906, and he corrected a small mistake in A. Einstein, Berichtigung zu meiner Arbeit: 'Eine neue Bestimmung der Moleküldimensionen', *Annalen der Physik* **34**, pp. 591–592, 1911, where, using new data, he found the value $6.6 \cdot 10^{23}$ for Avogadro's number. Cited on page 221.

173 The first tests of the prediction were performed by J. Perrin, *Comptes Rendus de l'Académie des Sciences* **147**, pp. 475–476, and pp. 530–532, 1908. He masterfully sums up the whole discussion in Jean Perrin, *Les atomes*, Librarie Félix Alcan, Paris, 1913. Cited on page 222.

174 Pierre Gaspard & al., Experimental evidence for microscopic chaos, *Nature* **394**, p. 865, 27 August 1998. Cited on page 223.

175 Gas expansion was the main method used for the definition of the official temperature scale. Only in 1990 were other methods introduced officially, such as total radiation thermometry (in the range 140 K to 373 K), noise thermometry (2 K to 4 K and 900 K to 1235 K), acoustical thermometry (around 303 K), magnetic thermometry (0.5 K to 2.6 K) and optical radiation thermometry (above 730 K). Radiation thermometry is still the central method in the range from about 3 K to about 1000 K. This is explained in detail in R.L. Rusby, R.P. Hudson, M. Durieux, J.F. Schooley, P.P.M. Steur & C.A. Swenson, The basis of the ITS-90, *Metrologia* **28**, pp. 9–18, 1991. On the water boiling point see also page 1076. Cited on page 214.

176 See for example the captivating text by Gino Segrè, *A Matter of Degrees: What Temperature Reveals About the Past and Future of Our Species, Planet and Universe*, Viking, New York, 2002. Cited on page 214.

177 See for example the article by H. Preston-Thomas, The international temperature scale of 1990 (ITS-90), *Metrologia* **27**, pp. 3–10, 1990, and the errata H. Preston-Thomas, The international temperature scale of 1990 (ITS-90), *Metrologia* **27**, p. 107, 1990, Cited on page 221.

178 For an overview, see Christian Enss & Siegfried Hunklinger, *Low-Temperature Physics*, Springer, 2005. Cited on page 221.

179 The quote is found in the introduction by George Wald to the text by Lawrence J. Henderson, *The Fitness of the Environment*, Macmillan, New York, 1913, reprinted 1958. Cited on page 226.

180 A fascinating introduction to chemistry is the text by John Emsley, *Molecules at an Exhibition*, Oxford University Press, 1998. Cited on page 227.

181 An excellent introduction into the physics of heat is the book by Linda Reichl, *A Modern Course in Statistical Physics*, Wiley, 2nd edition, 1998. Cited on page 228.

182 See V.L. Telegdi, Enrico Fermi in America, *Physics Today* **55**, pp. 38–43, June 2002. Cited on page 228.

183 K. Schmidt-Nielsen, *Desert Animals: Physiological Problems of Heat and Water*, Oxford University Press, 1964. Cited on page 229.

184 Why entropy is created when information is erased, but not when it is acquired, is explained in Charles H. Bennett & Rolf Landauer, Fundamental Limits of Computation, *Scientific American* **253:1**, pp. 48–56, 1985. The conclusion: we should pay to throw the newspaper away, not to buy it. Cited on page 230.

185 See, for example, G. Swift, Thermoacoustic engines and refrigerators, *Physics Today* **48**, pp. 22–28, July 1995. Cited on page 231.

186 K. Mertens, V. Putkaradze & P. Vorobieff, Braiding patterns on an inclined plane, *Nature* **430**, p. 165, 2004. Cited on page 238.

187 G. Müller, Starch columns: analog model for basalt columns, *Journal of Geophysical Research* **103**, pp. 15239–15253, 1998. Cited on page 239.

188 F. Herrmann, Mengenartige Größen im Physikunterricht, *Physikalische Blätter* **54**, pp. 830–832, September 1998. See also his lecture notes on general introductory physics on the website http://www-tfp.physik.uni-karlsruhe.de/~didaktik. Cited on pages 213 and 217.

189 B. Polster, What is the best way to lace your shoes?, *Nature* **420**, p. 476, 5 December 2002. Cited on page 218.

190 These points are made clearly and forcibly, as is his style, by N.G. van Kampen, Entropie, *nederlands tijdschrift voor natuurkunde* **62**, pp. 395–396, 3 December 1996. Cited on page 224.

191 This is a disappointing result of all efforts so far, as Grégoire Nicolis always stresses in his university courses. Seth Lloyd has compiled a list of 31 proposed definitions of complexity, containing among others, fractal dimension, grammatical complexity, computational complexity, thermodynamic depth. See, for example, a short summary in *Scientific American* p. 77, June 1995. Cited on page 224.

192 Minimal entropy is discussed by L. Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, *Zeitschrift für Physik* **53**, pp. 840–856, 1929. This classic paper can also be found in English translation in his collected works. Cited on page 225.

193 G. Cohen-Tannoudji, *Les constantes universelles*, Pluriel, Hachette, 1998. See also L. Brillouin, *Science and Information Theory*, Academic Press, 1962. Cited on page 225.

194 See for example A.E. Shalyt-Margolin & A.Ya. Tregubovich, Generalized uncertainty relation in thermodynamics, http://www.arxiv.org/abs/gr-qc/0307018, or J. Uffink & J. van Lith-van Dis, Thermodynamic uncertainty relations, *Foundations of Physics* **29**, p. 655, 1999. Cited on page 225.

195 B. Lavenda, *Statistical Physics: A Probabilistic Approach*, Wiley-Interscience, New York, 1991. Cited on pages 225 and 226.

196 H.W. Zimmermann, Particle entropies and entropy quanta IV: the ideal gas, the second law of thermodynamics, and the P-t uncertainty relation, *Zeitschrift für physikalische Chemie* **217**, pp. 55–78, 2003, and H.W. Zimmermann, Particle entropies and entropy quanta V: the P-t uncertainty relation, *Zeitschrift für physikalische Chemie* **217**, pp. 1097–1108, 2003. Cited on pages 225 and 226.

197 Quoted in D. Campbell, J. Crutchfield, J. Farmer & E. Jen, Experimental mathematics: the role of computation in nonlinear science, *Communications of the Association of Computing Machinery* **28**, pp. 374–384, 1985. Cited on page 232.

198 For more about the shapes of snowflakes, see the famous book by W.A. Bentlye & W.J. Humphreys, *Snow Crystals*, Dover Publications, New York, 1962. This second printing of the original from 1931 shows a large part of the result of Bentley's lifelong passion, namely several thousand photographs of snowflakes. Cited on page 232.

199 K. Schwenk, Why snakes have forked tongues, *Science* **263**, pp. 1573–1577, 1994. Cited on page 233.

200 P.B. Umbanhowar, F. Melo & H.L. Swinney, Localized excitations in a vertically vibrated granular layer, *Nature* **382**, pp. 793–796, 29 August 1996. Cited on page 233.

201 D.K. Campbell, S. Flach & Y.S. Kivshar, Localizing energy through nonlinearity and discreteness, *Physics Today* **57**, pp. 43–49, January 2004. Cited on page 233.

202 B. Andreotti, The song of dunes as a wave-particle mode locking, *Physical Review Letters* **92**, p. 238001, 2004. Cited on page 234.

203 K. Kötter, E. Goles & M. Markus, Shell structures with 'magic numbers' of spheres in a swirled disk, *Physical Review E* **60**, pp. 7182–7185, 1999. Cited on page 234.

204 A good introduction is the text by Daniel Walgraef, *Spatiotemporal Pattern Formation, With Examples in Physics, Chemistry and Materials Science*, Springer 1996. Cited on page 234.

205 For an overview, see the Ph.D. thesis by Joceline Lega, *Défauts topologiques associés à la brisure de l'invariance de translation dans le temps*, Université de Nice, 1989. Cited on page 235.

206 For a short, modern overview of turbulence, see L.P. Kadanoff, A model of turbulence, *Physics Today* **48**, pp. 11–13, September 1995. Cited on page 237.

207 For a clear introduction, see T. Schmidt & M. Mahrl, A simple mathematical model of a dripping tap, *European Journal of Physics* **18**, pp. 377–383, 1997. Cited on page 237.

208 An overview of science humour can be found in the famous anthology compiled by R.L. Weber, edited by E. Mendoza, *A Random Walk in Science*, Institute of Physics, 1973. It is also available in several expanded translations. Cited on page 237.

209 B. Hof, C.W.H. van Doorne, J. Westerweel, F.T.M. Nieuwstadt, H. Wedin, R. Kerswell, F. Waleffe, H. Faisst & B. Eckhardt, Experimental observation of nonlinear traveling waves in turbulent pipe flow, *Science* **305**, pp. 1594–1598, 2004. Cited on page 239.

210 J.J. Lissauer, Chaotic motion in the solar system, *Reviews of Modern Physics* **71**, pp. 835–845, 1999. Cited on page 239.

211 The present state of our the understanding of turbulence is described in ... Cited on page 240.

212 See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, 1991, p. 426. Cited on page 240.

CHAPTER II

SPECIAL RELATIVITY

The freedom of motion that is allowed by the Galilean description is only apparent. The first limitation we discover is the existence of a maximal speed in nature. Like all limits, also this one produces many fascinating results. The maximum speed leads to observer-varying time and length intervals, to an intimate relation between mass and energy, and to the existence of event horizons.

6. MAXIMUM SPEED, OBSERVERS AT REST, AND MOTION OF LIGHT

Fama nihil est celerius.*

LIGHT is indispensable for describing motion with precision. Checking whether a line or a path of motion is straight requires to look along it. In other words, we use light to define straightness. How do we decide whether a plane is flat? We look across it,** again using light. How do we measure length to high precision? With light. How do we measure time to high precision? With light; once that from the Sun was used, nowadays it is light from caesium atoms.*** In other words, light is important because it is the official standard for *undisturbed motion*. Physics would have evolved much more rapidly if, at some earlier time, light propagation had been recognized as the ideal example of motion.

But is light a moving phenomenon at all? It was already known in ancient Greece that this can be proven by a simple daily phenomenon, the *shadow*. Shadows prove that light is a moving entity, emanating from the light source, and moving in straight lines.**** The obvious conclusion that light takes a certain amount of time to travel from the source to Ref. 213 the surface showing the shadow had already been reached by the Greek thinker Empedocles (*c.* 490 to *c.* 430 BCE).

* 'Nothing is faster than rumor.' The sentence is a simplified version of the phrase by Vergil: fama, malum qua non aliud velocius ullum. 'Rumor, the evil faster than all.' From the *Aeneis*, book IV, verse 173 and 174.

** Note that looking along the plane from all sides is not sufficient for this; a surface that a light beam touches right along its length in *all* directions does not need to be flat. Can you give an example? One needs Challenge 526 n other methods to check flatness with light. Can you specify one?

*** For more details on the definition of length, see Appendix B on physical measurement units, on page 1060.

**** Whenever a source produces shadows, the emitted entities are called *rays* or *radiation*. Apart from light, other examples of radiation discovered through shadows were *infrared rays* and *ultraviolet rays*, which emanate from most light sources together with visible light, and *cathode rays*, which were found to be to the motion of a new particle, the *electron*. Shadows also led to the discovery of *X-rays*, which again turned out to be a – high frequency – version of light. Also *channel rays* showed up via their shadows; channel rays turn out to be travelling ionized atoms. The three types of radioactivity, namely α*-rays* (helium nuclei), β*-rays* (again electrons), and γ*-rays* (high energy X-rays) also produce shadows. All these discoveries were made between 1890 and 1910; those were the 'ray days' of physics.

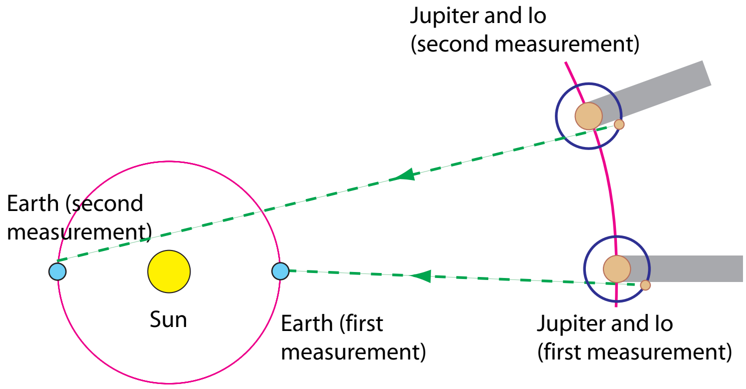


Figure 127 Rømer's method of measuring the speed of light

We can confirm this result with a different, but equally simple, argument. Speed can be measured. Therefore the *perfect* speed, which is used as the implicit measurement standard, must have a finite value. An infinite velocity standard would not allow measurements at all. In nature, the lightest entities move with the highest speed. Light, which is everything but heavy, is an obvious candidate for motion with perfect but finite speed. We will confirm this in a minute.

Challenge 527 n

A finite speed of light means that whatever we see is a message from the *past*. When we see the stars, the Sun or a loved one, we always see an image of the past. In a sense, experiments show us that nature prevents us from enjoying the present – we must therefore learn to enjoy the past.

The speed of light is high; therefore it was not measured for the first time until 1676, even though many, including Galileo, had tried to do so earlier. The first measurement method was discovered by the Danish astronomer Olaf Rømer (1644–1710) when he studied the orbits of Io and the other moons of Jupiter. He obtained an incorrect value for the speed of light because he used the wrong value for their distance from Earth. However, this was quickly corrected by his peers, including Newton himself. You might try to deduce his method from Figure 127. Since that time it is known that light takes a bit more than 8 minutes from the Sun to the Earth. The result was confirmed most beautifully by the next measurement, which was performed only fifty years later, in 1726, by the astronomer James Bradley (1693–1762). Being English, Bradley thought of the 'rain method' to measure the speed of light.

Challenge 528 n

Ref. 214

How can we measure the speed of falling rain? We walk rapidly with an umbrella, measure the angle α at which the rain appears to fall, and then measure our own velocity v. As shown in Figure 128, the velocity c of the rain is then given by

$$c = v / \tan\alpha \; . \qquad (100)$$

The same measurement can be made for light; we just need to measure the angle at which the light from a star above Earth's orbit arrives at the Earth. This effect is called the *aberration* of light; the angle is found most easily by comparing measurements distant by six months. The value of the angle is 20.5″; nowadays it can be measured with

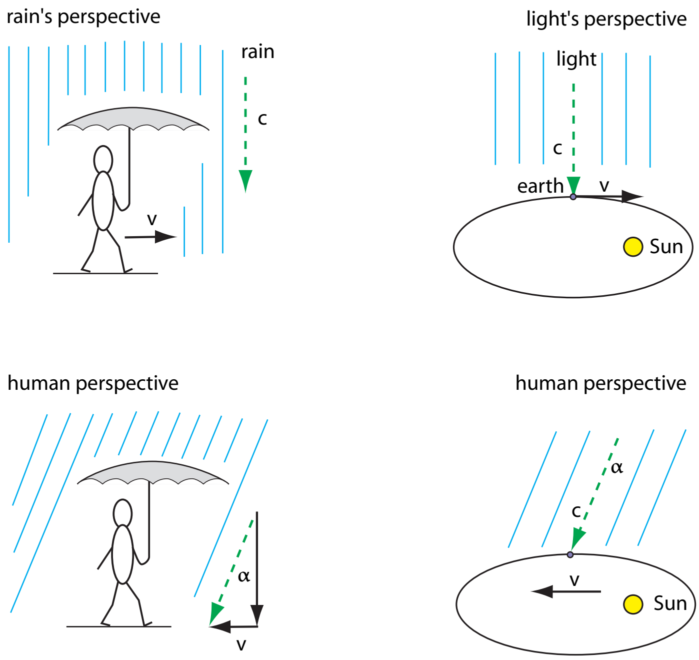


Figure 128 The rain method of measuring the speed of light

a precision of five decimal digits. Given that the velocity of the Earth around the Sun is $v = 2\pi R/T = 29.7\,\mathrm{km/s}$, the speed of light must therefore be $c = 3.00 \cdot 10^8\,\mathrm{m/s}$.* This

* Umbrellas were not common in Britain in 1726; they became fashionable later, after being introduced from China. The umbrella part of the story is made up. In reality, Bradley first understood his unexpected result while sailing on the Thames, when he noted that on a moving ship the apparent wind has a different direction to that on land. He had observed 50 stars for many years, notably al gamma Draconis, and during that time he had been puzzled by the *sign* of the aberration, which was *opposite* to the effect he was looking for, namely the star parallax. Both the parallax and the aberration for a star above the eclipse makes them describe a small ellipse in the course of the 12 month of a year, though with different rotation sense. Can
Challenge 529 n you see why?

By the way, it follows from special relativity that the formula (100) is wrong, and that the correct formula
Challenge 530 n is $c = v/\sin a$; can you see why?

To determine the velocity of the Earth, its distance to the Sun has to be determined. This is done most simply by a method published already by the Greek thinker Aristarchos of Samos (*c.* 310 to *c.* 230 BCE). You measure the angle between the Moon and the Sun at the moment that the Moon is precisely half full. The cosine of that angle gives the ratio between the distance to the Moon (determined e.g. via the methods of
Challenge 531 n page 109) and the distance to the Sun. The explanation is a puzzle left to the reader.

The angle in question is almost a right angle (which would yield an infinite distance), and good instru-
Ref. 215 ments are needed to measure it with precision, as Hipparchos noted in an extensive discussion of the problem around 130 BCE. The measurement became possible only in the late seventeenth century, showing that
Page 1071 its value is 89.86°, and the distance ratio about 400. Today, through radar measurements of planets, the distance to the Sun is known with the incredible precision of 30 metres. Moon distance variations can even be
Challenge 532 n measured down to the 1 centimetre range; can you guess how this is achieved?
Ref. 63 Aristarchos also determined the radius of the Sun and of the Moon as multiples of those of the Earth. Aristarchos was a remarkable thinker: he was the first to propose the heliocentric system, and perhaps the first to propose that stars were other, far away suns. For these ideas, several contemporaries of Aristarchos

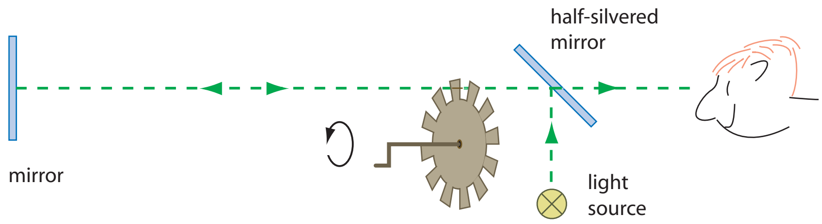


Figure 129 Fizeau's set-up to measure the speed of light

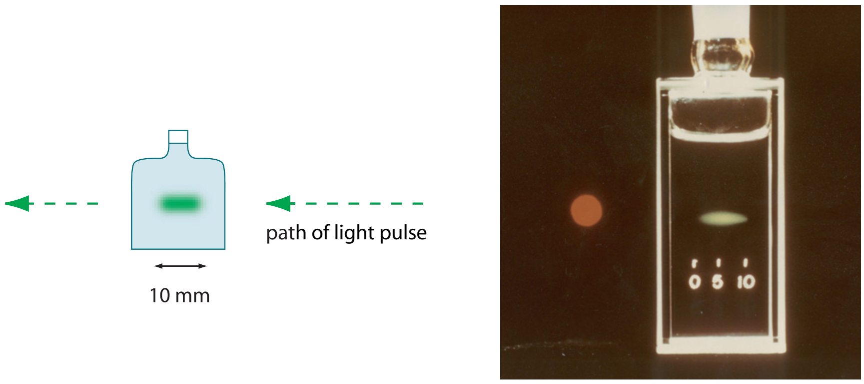


Figure 130 A photograph of a light pulse moving from right to left through a bottle with milky water, marked in millimetres (© Tom Mattick)

is quite an astonishing value, especially when compared with the fastest velocity ever achieved by a man made object, namely the Voyager satellites, which travel at 52 Mm/h = 14 km/s, with the growth of children, about 3 nm/s, or with the growth of stalagmites in caves, about 0.3 pm/s. We begin to realize why the speed of light measurements are a science of its own.

The first *precise* measurement of the speed of light was performed in 1849 by the French physicist Hippolyte L. Fizeau (1819–1896). His value was only 5 % greater than the modern one. He sent a beam of light towards a distant mirror and measured the time the light took to come back. How far away does the mirror have to be? How did Fizeau measure the time without any electric device? Part of the answer is given in Figure 129. Today, the experiment is much simpler; in the chapter on electrodynamics we will discover how to measure the speed of light using two standard UNIX or Linux computers connected by a cable.

Challenge 533 n

Page 517

The speed of light is so high that it is even difficult to prove that it is *finite*. Perhaps the most beautiful way to prove this is to photograph a light pulse flying across one s field of view, in the same way as one takes the picture of a car driving by or of a bullet flying along. Figure 130 shows the first such photograph, produced in 1971 with a standard off-the-shelf

Ref. 216

proposed that he should be condemned to death for impiety. When the Polish monk and astronomer Nicolaus Copernicus (1473–1543) reproposed the heliocentric system two thousand years later, he kept this reference unmentioned, even though he got the idea from him.

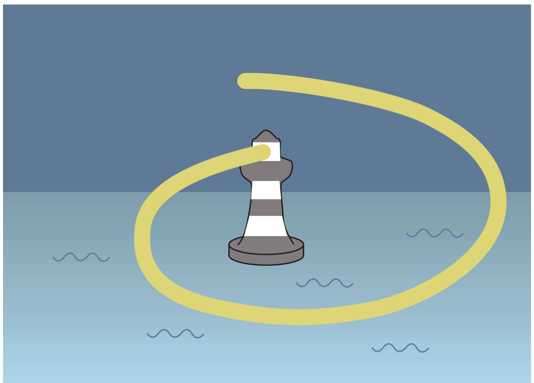


Figure 131 A consequence of the finiteness of the speed of light

Table 33 Properties of the motion of light

Observations about light
light can move through vacuum;
light transports energy;
light has momentum: it can hit bodies;
light has angular momentum: it can rotate bodies;
light moves across other light undisturbed;
light in vacuum always moves faster than any material body does;
the speed of light, its true signal speed, is the forerunner speed;
in vacuum its value is 299 792 458 m/s;
the proper speed of light is infinite;
shadows can move without any speed limit;
light moves straight when far from matter;
high intensity light is a wave;
light beams are approximations when wavelength is neglected;
in matter, both the forerunner speed and the energy speed of light are lower than in vacuum;
in matter, the group velocity of light pulses can be zero, positive, negative or infinite.

reflex camera, a very fast shutter invented by the photographers, and, most noteworthy, not a single piece of electronic equipment. (How fast does such a shutter have to be? How would you build such a shutter? And how would you open it at the right instant?) Challenge 534 n

A finite speed also implies that a rapidly rotating light beam behaves as shown as in Figure 131. In everyday life, the large velocity of light and the slow rotation velocity of lighthouses make the effect barely noticeable.

In short, light moves extremely rapidly. It is thus much faster than lightning, as you might like to check yourself. A century of precision measurements of the speed can be summarized in one result: Challenge 535 n

$$c = 299\,792\,458\ \mathrm{m/s}. \qquad (101)$$

Nowadays the measurement is so precise and easy, that the value has been fixed exactly, by redefining the meter appropriately. Table 33 gives a summary about what is known today about the motion of light. Once the velocity of light could be measured routinely, two surprising properties were discovered in the late nineteenth century. They form the basis of special relativity. Ref. 217

Can one play tennis using a laser pulse as the ball and mirrors as rackets?

Et nihil est celerius annis.*

Ovidius, *Metamorphoses*.

We all know that in order to throw a stone as far as possible, we run as we throw it; we know instinctively that in that case the stone's speed with respect to the ground is higher. However, to the initial astonishment of everybody, experiments show that light emitted from a moving lamp has the same speed as light emitted from a resting one. Light (in vacuum) is never faster than light; all light beams have the same speed. Many carefully and specially designed experiments confirmed this result to high precision; the speed of light can be measured with a precision of better than 1 m/s, but even for lamp speeds of more than 290 000 000 m/s no differences have been found. (Can you guess what lamps were used?) In everyday life, we know that a stone arrives more rapidly if we run towards it. Again, for light no difference is measured. All experiments show that the velocity of light has the *same value* for all observers, even if they are moving with respect to each other or with respect to the light source. The velocity of light is indeed the ideal, perfect measurement standard.** Ref. 218 Challenge 536 n

There is also a second set of experimental evidence for the constancy of the speed of light: every electromagnetic device, such as an electric toothbrush, shows that the speed of light is constant. We will discover that magnetic fields would not result from electric currents, as they do every day in every motor and in every loudspeaker, if the speed of light were not constant. This was actually the historical way the constancy was first deduced by several researchers. Only after realizing this connection, did the German–Swiss physicist Albert Einstein*** show that the constancy is also in agreement with the motion of bodies, as we will do in this section. The connection between electric toothbrushes Ref. 221 Page 497

* 'Nothing is faster than the years.' Book X, verse 520.

** An equivalent alternative term for the speed of light is 'radar speed' or 'radio speed'; we will see below why this is the case.

The speed of light is also not far from the speed of neutrinos. This was shown most spectacularly by the observation of a supernova in 1987, when the flash and the neutrino pulse arrived spaced by a few hours. Can you deduce the maximal difference between the two speeds, knowing that the supernova was $1.7 \cdot 10^5$ light years away? Challenge 537 n

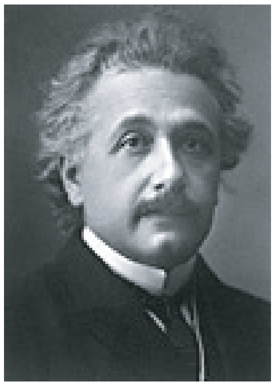
Experiments also show that the speed of light is the same in all directions of space to at least 21 digits of precision. Other data, taken from gamma ray bursts, show that the speed of light is independent of frequency for its first 20 digits at least. Ref. 219 Ref. 220

*** Albert Einstein (b. 1879 Ulm, d. 1955 Princeton); one of the greatest physicists. He published three important papers in 1905, namely about Brownian motion, about special relativity and about the idea of light quanta. Each paper was worth a Nobel prize, but he was awarded the prize only for the last one. In 1905, he also discovered the famous formula $E_0 = mc^2$ (published early 1906). Although he was one of the founders of quantum theory, he later turned against it. His famous discussions with his friend Niels Bohr nevertheless helped to clarify the field in its most counter-intuitive aspects. He explained the Einstein-de Haas effect that proves that magnetism is due to motion inside materials. In 1915 and 1916, he published the and relativity will be detailed in the chapter on electrodynamics.* In simple words, if the speed of light were not constant, observers would be able to move at the speed of light. Since light is a wave, such observers would see a wave standing still. However, electromagnetism forbids the existence of such a phenomenon. Therefore, observers cannot reach the speed of light. In summary, the velocity v of any physical system in nature (i.e. any set of localized energy) is bound by

$$v \leqslant c \tag{102}$$

This relation is the basis of special relativity; in fact, the full theory of special relativity is contained in it. Einstein often regretted that the theory was called 'Relativitätstheorie' or 'theory of relativity'; he preferred the name 'Invarianztheorie' or 'theory of invariance', but was not able to change the name. Ref. 224

The constancy of the speed of light is in complete contrast with Galilean mechanics, and proves that the latter is *wrong* at high velocities. At low velocities the description remains good, because the error is small. But if we look for a description valid at *all* velocities, Galilean mechanics has to be discarded. For example, when we play tennis we use the observation that by hitting the ball in the right way, we can increase or decrease its speed. But with light this is impossible. Even if we take an aeroplane and fly after a light beam, it still moves away with the same speed. Light does not behave like cars. If we accelerate a bus we are driving, the cars on the other side of the road pass by with higher and higher speeds as we drive faster. For light, this is *not* so; light always passes by with the *same* speed.*



Albert Einstein

Why is this result almost unbelievable, even though the measurements show it unambiguously? Take two observers O and Ω (pronounced 'omega') moving with relative velocity v, such as two cars on opposite sides of the street. Imagine that at the moment they pass each other, a light flash is emitted by a lamp in the hand of O. The light flash moves through positions $x(t)$ for O and through positions $\xi(\tau)$ (pronounced 'xi of tau') for Ω. Since the speed of light is the same for both, we have

$$\frac{x}{t} = c = \frac{\xi}{\tau} \,. \tag{103}$$

general theory of relativity, one of the most beautiful and remarkable works of science ever. Being Jewish and famous, he was a favourite target of attacks and discrimination by the national-socialist movement; in 1933 he emigrated to the USA. He was not only a great physicist, but also a great thinker; reading his collection of thoughts about topics outside physics is time well spent. Ref. 222

All those interested in emulating Einstein should know that he published many papers, and that many of them were wrong; he then corrected the calculations results in subsequent papers, and then again. This happened so frequently that he made fun of himself about this. This reminds one of the famous definition that a genius is a person that makes the largest number of mistakes possible in the shortest lapse of time possible.

* For information about the influences of relativity on machine design see the interesting textbook by Van Bladel. Ref. 223

* Indeed, the presently possible measurement precision of $2 \cdot 10^{-13}$ does not allow to discern any changes of the speed of light with the speed of the observer. Ref. 219

However, in the situation described, we obviously have $x \neq \xi$. In other words, the constancy of speed of light implies that $t \neq \tau$, i.e. that *time is different for observers moving relative to each other*. Time is thus not unique. This surprising result, which in the mean time has been confirmed by many experiments, was first stated in detail in 1905 by Albert Einstein. Though many had the data on the invariance of c on their desks, only the young Einstein had the courage to make the statement that time is observer dependent, and to face the consequences. Let us do it as well.

Challenge 538 e
Ref. 225

Already in 1895, the discussion of viewpoint invariance had been called the *theory of relativity* by Henri Poincaré.* Einstein called the description of motion without gravity the theory of *special relativity*, and the description with gravity the theory of *general relativity*. Both fields are full of fascinating and counter-intuitive results. In particular, they show that everyday, Galilean physics is wrong at high speeds, because there is a speed limit in nature.

Ref. 221

Special relativity in a few lines

The speed of light is constant for all observers. We thus can deduce all relations between what two different observers measure with the help of Figure 132. It shows two observers moving with constant speed against each other in space-time, with the first sending a light flash to the second, from where it is reflected back to the first. Since light speed is constant, light is the only way to compare time and space coordinates for two distant observers. Two distant clocks (like two distant meter bars) can only be compared, or synchronized, using light or radio flashes. Since light speed is constant, light paths are parallel in such diagrams.

Ref. 227

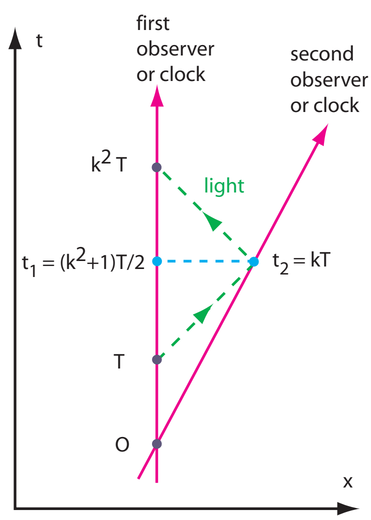


Figure 132 A drawing containing most of special relativity

A constant relative speed between two observers implies that a common factor k appears and relates the time coordinates of events. (Why is the relation linear?) If a flash starts at a time T as measured for the first observer, it arrives at the second at kT, and then back again to the first at time k^2T. The drawing shows that

Challenge 539 ny
Challenge 540 n

$$k = \sqrt{\frac{c+v}{c-v}} \quad \text{or} \quad \frac{v}{c} = \frac{k^2-1}{k^2+1} . \tag{104}$$

This factor will appear again in the Doppler effect.**

* Henri Poincaré (1854–1912), important French mathematician and physicist. Poincaré was one of the most productive men of his time, advancing relativity, quantum theory, and many part of mathematics.

The most beautiful and simple introduction to relativity is still given by Albert Einstein himself, such as in *Über die spezielle und allgemeine Relativitätstheorie*, Vieweg, 1997, or in *The Meaning of Relativity*, Methuen, London, 1951. Only a century later there are books almost as beautiful, such as the text by Taylor and Wheeler.

Ref. 226

** Explaining relativity with the factor k is often called *k-calculus*.

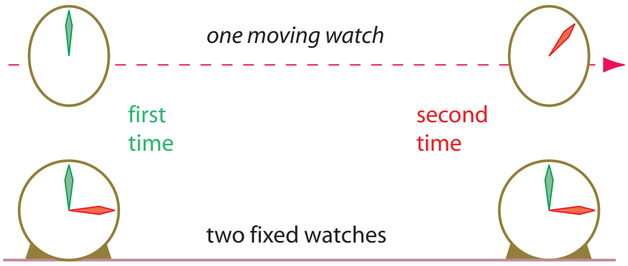


Figure 133 Moving clocks go slow

The figure also shows that the time coordinate t_1 assigned by the first observer to the moment in which the light is reflected is different from the coordinate t_2 assigned by the second observer. Time is indeed different for two observers in relative motion.

The *time dilation factor* between the two time coordinates is found from Figure 133 by comparing the values t_1 and t_2; it is given by

$$\frac{t_1}{t_2} = \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma(v) . \tag{105}$$

Time intervals for a moving observer are *shorter* by this factor γ; the time dilation factor is always larger than 1. In other words, *moving clocks go slower*. For everyday speeds the amount is tiny. That is why we do not detect time differences in everyday life. Nevertheless, Galilean physics is not correct for speeds near that of light. The same factor γ also appears in the formula $E = \gamma mc^2$ that we will deduce below. Expression (104) or (105) is the only piece of mathematics needed in special relativity; all other results derive from it.

If a light flash is sent forward and back starting from the second observer, he will make the same statement: for him, the first clock is moving, and also for him, the moving clock goes slower. The situation is similar to a man comparing the number of steps between two identical ladders that are not parallel. A man on either ladder will always observe that the steps of the *other* ladder are shorter. Alternatively, take two people moving away from each other: each of them notes that the other gets smaller as their distance increases.

Obviously, many people tried to find arguments to avoid the strange conclusion that time differs from observer to observer. But all had to bow to the experimental results. Let us have a look at some of them.

Acceleration of light and the Doppler effect

Light *can* be accelerated. Every mirror does this! We will see in the chapter on electromagnetism that matter also has the power to *bend* light, and thus to accelerate it. However, it
Page 524
will turn out that all these methods only change the propagation direction; none has the power to change the speed of light in a vacuum. In short, light is an example of motion
Challenge 541 n
which cannot be stopped. Only a few other examples exist. Can you name one?

What would happen if we could accelerate light to higher speeds? In that case light would be made of particles with non-vanishing mass. Physicists call such particles

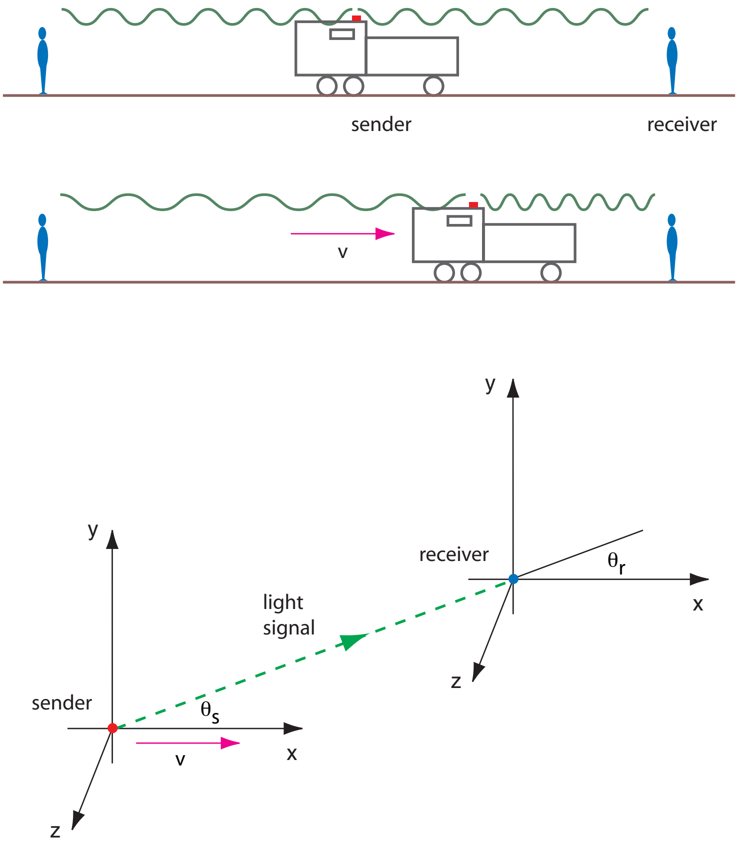


Figure 134 The set-up for the observation of the Doppler effect

massive particles. If light had mass, it would be necessary to distinguish the 'massless energy speed' c from the speed of light c_l, which then would be lower and depend on the kinetic energy of those massive particles. The speed of light would not be constant, but the massless energy speed would still be so. Massive light particles could be captured, stopped and stored in a box. Such boxes would render electric illumination superfluous; it would be sufficient to store in them some daylight and release the light, slowly, the following night, maybe after giving it an additional push to speed it up.*

Ref. 228, Ref. 229 Physicists have therefore tested the possibility of massive light in quite some detail. Observations now put any possible mass of light (particles) at less than $1.3 \cdot 10^{-52}$ kg from terrestrial arguments, and at less than $4 \cdot 10^{-62}$ kg from astrophysical arguments. In other words, light is not heavy, light is light.

But what happens when light hits a *moving* mirror? If the speed of light does not change, something else must. The situation is akin to a light source moving with respect to the receiver; the receiver will observe a *different colour* from that observed by the sender.

* We mention for completeness that massive light would also have *longitudinal* polarization modes, also in contrast to observations, which show that light is polarized exclusively *transversally* to the propagation direction.

This result is called the *Doppler effect*. Christian Doppler was the first to study the frequency shift in the case of sound waves – the well-known change in whistle tone between approaching and departing trains – and to extend the concept to the case of light waves.* As we will see later on, light is (also) a wave, and its colour is determined by its frequency, or equivalently, by its wavelength λ. Like the tone change for moving trains, Doppler followed that a moving light source produces a colour at the receiver that is different from colour at the sending source. Simple geometry, and the conservation of the number of maxima and minima, leads to the result

Challenge 542 e

$$\frac{\lambda_R}{\lambda_S} = \frac{1}{\sqrt{1 - v^2/c^2}}(1 - \frac{v}{c}\cos\theta_R) = \gamma\,(1 - \frac{v}{c}\cos\theta_R)\ . \qquad (106)$$

The variables in this expression are defined in Figure 134. Light from an approaching source is thus blue shifted, whereas light from a departing source is red-shifted. The first observation of the Doppler effect for light was made by Johannes Stark* in 1905, by studying the light emitted by moving atoms. All subsequent experiments confirmed the calculated colour shift within measurement errors; the latest checks found agreement to within two parts per million. In contrast to sound waves, a colour change is also found when the motion is *transverse* to the light signal. Thus, a yellow rod in rapid motion across the field of view will have a blue leading edge and a red trailing edge prior to the closest approach to the observer. The colours result from a combination of the longitudinal (first-order) Doppler shift and the transverse (second-order) Doppler shift. At a particular angle $\theta_{\text{unshifted}}$ the colour will be the same. (How does the wavelength change in the purely transverse case? What is the expression for $\theta_{\text{unshifted}}$ in terms of v?)

Ref. 230

Challenge 543 n

The colour shift is used in many applications. Almost all solid bodies are mirrors for radio waves. When one enters a building, often the doors open automatically. A little sensor above the door detects the approaching person. Usually, but not always, this is done by measuring the Doppler effect of radio waves emitted by the sensor and reflected by the approaching person. (We will see later that radio waves and light are two sides of the same phenomenon.) In this way, doors open whenever something moves towards them. Police radar also uses the Doppler effect, this time to measure the speed of cars.**

The Doppler effect also makes it possible to measure the velocity of light sources. Indeed, it is commonly used to measure the speed of far away stars. In these cases, the Doppler shift is often characterized by the *red-shift number z*, defined with the help of

* Christian Andreas Doppler (b. 1803 Salzburg, d. 1853 Venezia), Austrian physicist. Doppler studied the effect named after him for sound and light. Already in 1842 he predicted that one day one could use the effect to measure the motion of distant stars by looking at their colours.

* Johannes Stark (1874–1957), discovered in 1905 the optical Doppler effect in channel rays, and in 1913 the splitting of spectral lines in electrical fields, nowadays called the Stark effect. For the two discoveries he received the 1919 Nobel prize for physics. He left his professorship in 1922 and later turned into a full-blown national socialist. Member of the NSDAP from 1930 onwards, he became known for aggressively criticizing other people's statements about nature purely for ideological reasons; he became rightly despised by the academic community all over the world.

Challenge 544 n

** At what speed does a red traffic light appear green?

wavelength λ or frequency F by

$$z = \frac{\Delta\lambda}{\lambda} = \frac{f_S}{f_R} - 1 = \sqrt{\frac{c+v}{c-v}} - 1 \,. \tag{107}$$

Challenge 545 n Can you imagine how the number z is determined? Typical values for z found for light sources in the sky range from -0.1 to 3.5, but higher values, up to more than 10, have also been found. Can you determine the corresponding speeds? How can they be so high? Challenge 546 n

In summary, whenever one tries to change the *speed* of light, one only manages to change its *colour*. That is the *Doppler effect*.

Page 123 We now from classical physics that when light passes a large mass, such as a star, it is deflected. Does this deflection lead to a Doppler shift? Challenge 547 n

The difference between light and sound

The Doppler effect for light is much more important than the Doppler effect for sound. Even if the speed of light were not yet known to be constant, the colour change alone already would *prove* that time is different for observers moving relative to each other. Why? Time is what we read from our watch. In order to determine whether another watch is synchronized with our own one, we look back and forward between the two. In short, we need to use light signals to synchronize clocks. Ref. 231 Now, any colour change when light moves from one observer to another necessarily implies that the watches run differently, and thus means that time is *different* at the two places. One way to deduce this is to note that also a light source is a clock – though one beating very rapidly. Given that two observers see different colours from the same source implies that they measure different oscillation or clock frequencies for the same clock. In other words, time is different for observers moving against each other. Indeed, relativity follows from the full Doppler effect for light. This is shown by equation (104). (Can you confirm that the connection between observer-dependent frequencies and observer-dependent time is not given when the Doppler effect for *sound* is used?) Challenge 548 n

Why does light imply special relativity, but sound in air does not? Light is a limit for the motion for energy, whereas sound in air is no such limit. Experience shows that there are supersonic aeroplanes, but there are no superluminal rockets. Observations thus show that the limit $v \leqslant c$ is valid only if c is the speed of light, not if c is the speed of sound in air.

However, there is one system in nature where the speed of sound is indeed a limit speed for energy: the speed of sound is the limit speed for the motion of *dislocations* in crystalline solids. (We discuss this in detail later on.) Page 910 As a result, the theory of special relativity is also valid for dislocations, provided that the speed of light is substituted everywhere by the speed of sound! Dislocations obey the Lorentz transformations, show length contraction and follow the famous energy formula $E = \gamma mc^2$. Ref. 232 In all these effects the speed of sound c plays the same role for dislocations that the speed of light plays for physical systems.

If special relativity is based on the statement that nothing can move faster than light, careful checks whether this is the case are required.

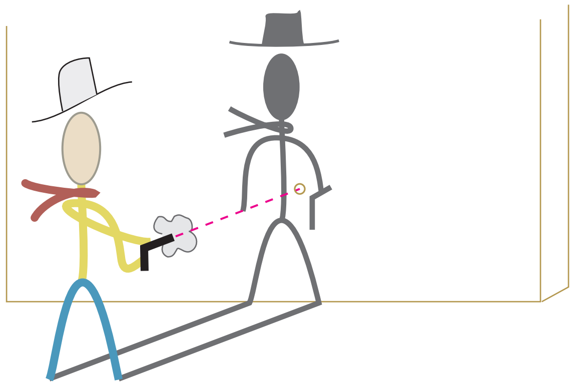


Figure 135 Lucky Luke

Can one shoot faster than one's shadow?

Quid celerius umbra?*

To realize what Lucky Luke does in Figure 135, the bullet has to move faster than the speed of light. (What about the hand?) To achieve this, certain people use the largest practical amounts of energy possible, taken directly from an electrical power station, accelerate the lightest bullets that can be handled, namely electrons, and measure the speed that results. This experiment is carried out daily in particle accelerators such as the Large Electron Positron ring, the LEP, of 27 km circumference located partly in France and partly in Switzerland, near Geneva. In that place, 40 MW of electrical power, the same amount used by a small city, accelerates electrons and positrons to energies of over 16 nJ (104.5 GeV) each. The result is shown in Figure 136: even with these impressive means it is impossible to make electrons move more rapidly than light. (Can you imagine a way to measure energy and speed separately?) The speed–energy relation of Figure 136 is a consequence of the maximum speed and is deduced below. These and many similar observations thus show that there is a *limit* to the velocity of objects. Velocities of bodies (or of radiation) higher than the speed of light do not exist.** Historically, the accuracy of Galilean mechanics was taken for granted for more than three centuries, so that nobody ever thought of checking it; but when this was finally done, as in Figure 136, it was found to be wrong.

Challenge 549 ny

Ref. 233

Challenge 550 ny

Page 285

The people most unhappy with this limit are computer engineers; if the speed limit were higher, it would be possible to make faster microprocessors and thus faster computers; this would allow, for example, more rapid progress towards the construction of computers that understand and use language.

* 'What is faster than the shadow?' A motto often found on sundials.

** There are still people who refuse to accept these results, as well as the ensuing theory of relativity. Every physicist should enjoy the experience, at least once in his life, of discussing with one of these men. (Strangely, no woman has yet been reported as member of this group of people.) This can be done e.g. via the internet, in the sci.physics.relativity news group. See also the http://www.crank.net website. Crackpots are a fascinating lot, especially since they teach the importance of *precision* in language and in reasoning, which they all, without exception, neglect. On the other hand, encounters with several of them provided the inspiration for this section.

Ref. 234

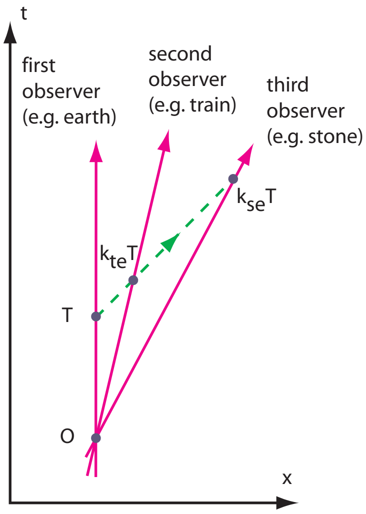


Figure 137 How to deduce the addition of velocities

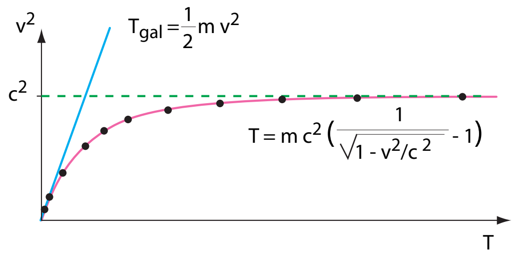


Figure 136 Experimental values (dots) for the electron velocity v as function of kinetic energy T

The observation of a limit speed is in complete contrast to Galilean mechanics. In fact, it means that for velocities near that of light, say about 15 000 km/s or more, the expression $mv^2/2$ is *not* equal to the kinetic energy T of the particle. In fact, such high speeds are rather common: many families have an example in their home. Just determine the speed of electrons inside a television, given that the transformer inside produces 30 kV.

Challenge 551 n

The observation of speed of light as a *limit* speed for objects is easily seen to be a consequence of its *constancy*. Bodies that can be at rest in one frame of reference obviously move more slowly than the maximum velocity (light) in that frame. Now, if something moves more slowly than something else for *one* observer, it does so for all other observers as well. (Trying to imagine a world in which this would not be so is interesting: funny things would happen, such as things interpenetrating each other.) Therefore no object that can be at rest can move faster than the limit speed. But any body which can be at rest does have different speeds for different observers. Conversely, if a phenomenon exists whose speed is the same for all observers, then this speed must necessarily be the limit speed. We also deduce that the maximum speed is the speed of *massless* entities. Light and all the other types of electromagnetic waves are the only known examples. The speed of gravitational waves is also predicted to achieve maximum speed. Though the speed of neutrinos cannot be distinguished experimentally from the maximum speed, recent experiments suggest that they do have a tiny mass.

Challenge 552 d

Challenge 553 e

Ref. 235

The addition of velocities

If the speed of light is a limit, all attempts to exceed it cannot lead to success. This implies that when speeds are composed, such as when a stone is thrown when running, the values cannot simply be added. If a train is travelling at velocity v_{te} compared to the Earth, and somebody throws a stone inside it with velocity v_{st} in the same direction, it is usually assumed as evident that the velocity of the stone relative to the Earth is given by $v_{se} = v_{st} + v_{te}$. In fact, both reasoning and measurement show a different result.

The existence of a maximum speed, together with Figure 137, implies that the k-factors must follow $k_{se} = k_{st} k_{te}$.* Then we only have to insert the relation (104) between each

Challenge 554 e

k-factor and the respective speed to get

$$v_{se} = \frac{v_{st} + v_{te}}{1 + v_{st} v_{te} / c^2} . \tag{108}$$

Challenge 555 e
Page 497

The result is never larger than c and is always smaller than the naive velocity sum of everyday life.** Expression (108) has been confirmed by literally all the millions of cases in which it has been checked so far.

Observers and the principle of special relativity

Special relativity is built on a simple principle:

▷ *The maximum speed of energy transport is the same for all observers.*

Ref. 237

Or, as Hendrik Lorentz*** liked to say, the equivalent:

▷ *The speed v of a physical system is bound by*

$$v \leqslant c \tag{109}$$

for all observers, where c is the speed of light.

Ref. 238

This independence of the speed of light from the observer was checked with high precision by Michelson and Morely**** in the years from 1887 onwards. It has been confirmed

* By taking the (natural) logarithm of this equation, one can define a quantity, the *rapidity*, that measures the speed and is additive.

Ref. 236

** One can also deduce the Lorentz transformation directly from this expression.

*** Hendrik Antoon Lorentz (b. 1853 Arnhem, d. 1928 Haarlem) was, together with Boltzmann and Kelvin, the most important physicist of his time. He deduced the so-called Lorentz transformation and the Lorentz contraction from Maxwell's equation of the electrodynamic field. He was the first to understand, long before quantum theory confirmed the idea, that Maxwell's equations for the vacuum describe also matter and all its properties, as long as moving charged point particles – the electrons – are included. He showed this in particular for the dispersion of light, for the Zeeman effect, for the Hall effect and for the Faraday effect. He gave the correct description of the Lorentz force. In 1902, he received the physics Nobel prize, together with Pieter Zeeman. Outside physics, he was active in the internationalization of scientific collaborations. He was also essential in the creation of the largest human-made structures on Earth: the polders of the Zuyder Zee.

**** Albert Abraham Michelson (b. 1852 Strelno, d. 1931 Pasadena) Prussian–Polish–US-American physicist, Nobel prize in physics in 1907. Michelson called the set-up he devised an *interferometer*, a term still in use today. Edward William Morely (1838–1923), US-American chemist, was Michelson's friend and long-time collaborator.

in all subsequent experiments. In fact, special relativity is also confirmed by all precision experiments performed *before* it was formulated. In addition, you can confirm it yourself Page 497 at home. These points are made in detail in the section on electrodynamics.

The existence of a limit speed allows to draw several interesting conclusions. To do this, we keep all other properties of Galilean physics intact.* The limit speed is the speed of light. It is constant for all observers. This constancy implies:

- In a closed free-floating room, there is no way to tell the speed of the room.
- There is no absolute rest; rest is an observer-dependent concept.**
- Time depends on the observer.

More interesting and specific conclusions can be drawn when two additional conditions are stated explicitly. First, we study situations where gravitation can be neglected. (If this not the case, we need *general* relativity to describe the system.) Second, we also assume that the data about the bodies under study – their speed, their position, etc. – can be gathered without disturbing them. (If this not the case, we need *quantum theory* to describe the system.)

To deduce the *precise* way that the different time intervals and lengths measured by two observers are related to each other, we take an additional simplifying step. We start with a situation where no interaction plays a role; in other words, we start with *relativistic kinematics* of bodies moving without disturbance.

If an undisturbed body travels along a straight line with a constant velocity, or if it stays at rest, one calls the observer making this observation *inertial*, and the coordinates used by the observer an *inertial frame of reference*. Every inertial observer is in undisturbed motion. Examples of inertial observers (or frames) are thus – for *two* dimensions – those moving on a frictionless ice surface or on the floor inside a smoothly running train or ship; a full example – for all *three* spatial dimensions – is a cosmonaut in an Apollo capsule while travelling between the Moon and the Earth, as long as the engine is switched off. Inertial observers in three dimensions might also be called *free-floating* observers. They are thus not so common. Non-inertial observers are much more common. Can you Challenge 557 e confirm this? Inertial observers are the most simple ones and form a special set:

- Any two inertial observers move with constant velocity relative to each other.
- All inertial observers are equivalent: they describe the world with the same equations. This statement was called the *principle of relativity* by Henri Poincaré. However, the *essence* of relativity is the existence of a limit speed.

To see how length and space intervals change from one observer to the other, we assume two inertial observers, a Roman one using coordinates x, y, z and t, and a Greek one using coordinates ξ, υ, ζ and τ,*** move with velocity $\mathbf{v}$ relative to each other. The

Page 81 * This point is essential. For example, Galilean physics states that only relative motion is physical. Galilean physics also excludes various mathematically possible options to realize a constant light speed that however would be in contradiction with everyday life.

Einstein's original 1905 paper starts from two principles: the constancy of the speed of light and the equivalence of all inertial observers. The equivalence of all inertial observers was already stated in 1632 by Galileo; only the constancy of the speed of light was new. Despite this fact, the new theory was named – by Poincaré – after the old principle.

Challenge 556 ny ** Can you confirm this deduction?

*** They are read as 'xi', 'upsilon', 'zeta' and 'tau'. The names, correspondences and pronunciations of all Greek letters are explained in Appendix A.

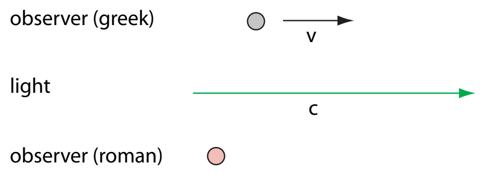


Figure 138 Two inertial observers, using coordinates (t, x) and (τ, ξ), and a beam of light

axes are chosen in such a way that the velocity points in the x-direction. The constancy of the speed of light in any direction for any two observers means that for the motion of light the coordinate differentials are related by

$$0 = (c\mathrm{d}t)^2 - (\mathrm{d}x)^2 - (\mathrm{d}y)^2 - (\mathrm{d}z)^2 = (c\mathrm{d}\tau)^2 - (\mathrm{d}\xi)^2 - (\mathrm{d}\upsilon)^2 - (\mathrm{d}\zeta)^2 . \quad (110)$$

Assume also that a flash lamp at rest for the Greek observer, thus with $\mathrm{d}\xi = 0$, produces two flashes spaced by an interval $\mathrm{d}\tau$. For the Roman observer, the flash lamp moves, so that $\mathrm{d}x = v\mathrm{d}t$. Inserting this into the previous expression, and assuming linearity and Challenge 558 e speed direction independence for the general case, we find that intervals are related by

$$\begin{aligned} \mathrm{d}t &= \gamma(\mathrm{d}\tau + v\mathrm{d}\xi/c^2) = \frac{\mathrm{d}\tau + v\mathrm{d}\xi/c^2}{\sqrt{1 - v^2/c^2}} \quad \text{with} \quad v = \mathrm{d}x/\mathrm{d}t \\ \mathrm{d}x &= \gamma(\mathrm{d}\xi + v\mathrm{d}\tau) = \frac{\mathrm{d}\xi + v\mathrm{d}\tau}{\sqrt{1 - v^2/c^2}} \\ \mathrm{d}y &= \mathrm{d}\upsilon \\ \mathrm{d}z &= \mathrm{d}\zeta \ . \end{aligned} \quad (111)$$

These expressions describe how length and time intervals measured by different observers are related. At relative speeds v that are small compared to the velocity of light, such as in everyday life, the time intervals are essentially equal; the *stretch factor* or *relativistic correction* or *relativistic contraction* γ is then equal to 1 for all practical purposes. However, for velocities *near* that of light the measurements of the two observers give different values. In these cases, space and time *mix*, as shown in Figure 139.

The expressions (111) are also strange in another respect. When two observers look Challenge 559 n at each other, each of them claims to measure shorter intervals than the other. In other words, special relativity shows that the grass on the other side of the fence is always *shorter* – if one rides along the fence on a bicycle and if the grass is inclined. We explore this bizarre result in more detail shortly.

The stretch factor γ is equal to 1 in everyday life and for most practical purposes. The largest value humans have ever achieved is about $2 \cdot 10^5$; the largest observed value in Challenge 560 n nature is about 10^{12}. Can you imagine their occurrences?

Once we know how space and time *intervals* change, we can easily deduce how *coordinates* change. Figures 138 and 139 show that the x coordinate of an event L is the sum of two intervals: the ξ coordinate and the length of the distance between the two origins.

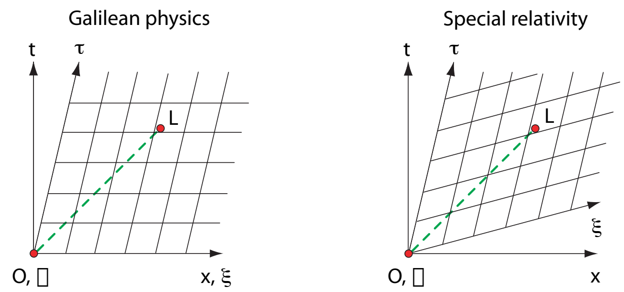


Figure 139 Space-time diagrams for light seen from two different observers

In other words, we have

$$\xi = \gamma(x - vt) \quad \text{and} \quad v = \frac{\mathrm{d}x}{\mathrm{d}t} \ . \tag{112}$$

Using the invariance of the space-time interval, we get

$$\tau = \gamma(t - xv/c^2) \ . \tag{113}$$

Henri Poincaré called these two relations the *Lorentz transformations of space and time* after their discoverer, the Dutch physicist Hendrik Antoon Lorentz.* In one of the most beautiful discoveries of physics, in 1892 and 1904, Lorentz deduced these relations from the equations of electrodynamics, which had contained them, waiting to be discovered, since 1865.** In that year James Clerk Maxwell had published the equations in order to describe everything electric and magnetic. However, only Einstein understood that t and τ, like x and ξ, are equally correct and thus equally valid descriptions of space and time.

Ref. 239

Page 504

The Lorentz transformation describes the change of viewpoint from one inertial frame to a second, moving one. This change of viewpoint is called a (Lorentz) *boost*. The formulae (112) and (113) for the boost are central to the theories of relativity, both the special and the general one. In fact, the mathematics of special relativity will not get more difficult than that; if you know what a square root is, you can study special relativity in all its beauty.

Many alternative formulae for boosts have been explored, such as expressions in which instead of the relative velocity also the relative acceleration of the two observers is included. However, all alternatives had to be discarded after comparing them to experimental results. Before we have a look at such experiments, we continue with a few logical deductions from the boost relations.

Ref. 240

What is space-time?

The Lorentz transformations tell something important: space and time are two aspects of the same 'stuff', they are two aspects of the same basic entity. They mix in different

* About Hendrik Antoon Lorentz, see page 263.

** The Irishman George F. Fitzgerald had had already discovered the Lorentz transformations in 1889, but had, in contrast to Lorentz, not continued his research in the field.

ways for different observers. This fact is commonly expressed by stating that time is the *fourth dimension*. This makes sense because the common entity – called *space-time* – can be defined as the set of all possible events, because events are described by four coordinates in time and space, and because the set of all events behaves like a manifold. (Can Challenge 561 n you confirm this?) In other words, the maximum speed in nature forces us to introduce space-time for the description of nature. In the theory of special relativity, the space-time manifold* is characterized by a simple property: the *space-time interval* di between two Ref. 241 nearby events, defined as

$$\mathrm{d}i^2 = c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 = c^2\mathrm{d}t^2\Big(1 - \frac{v^2}{c^2}\Big) , \qquad (114)$$

is *independent* of the (inertial) observer. Such a space-time is also called Minkowski space-time, after the German physicist Hermann Minkowski (1864–1909), the prematurely passed away teacher of Albert Einstein; he was the first, in 1904, to define the concept of space-time and to understand its usefulness and importance.

The space-time interval of equation (114) has a simple interpretation. It is the time measured by an observer moving from event (t, x) to event $(t + \mathrm{d}t, x + \mathrm{d}x)$, the so-called *proper time*, multiplied by c^2. We could simply call it wristwatch time.

How does Minkowski space-time differ from Galilean space-time, the combination of everyday space and time? Both space-times are manifolds, i.e. continuum sets of points, both have one temporal and three spatial dimensions, and both manifolds are infinite, Challenge 562 n i.e. open, with the topology of the punctured sphere. (Can you confirm this?) Both manifolds are flat, i.e. free of curvature. In both cases, space is what is measured with a metre rule or with a light ray, and time is what is read from a clock. In both cases, space-time is fundamental; it is and remains the *background* and the *container* of things and events. We *live* in a Minkowski space-time, so to speak. Minkowski space-time exists independently of things. And even though coordinate systems can be different from observer to observer, the underlying entity, space-time, is still *unique*, even though space and time by themselves are not.

The central difference, in fact the only one, is that Minkowski space-time, in contrast to the Galilean case, *mixes* space and time, and in particular, does so differently for observers with different speeds, as shown in Figure 139. That is the reason that time is an observer-dependent concept.

The maximum speed in nature thus forces us to describe motion with space-time. That is interesting, because in space-time, speaking in tabloid language, *motion does not exist*. Motion exists only in space. In space-time, nothing moves. For each point particle, space-time contains a world-line. In other words, instead of asking why motion exists, we can equivalently ask why space-time is criss-crossed by world-lines. At this point, we are still far from answering either question. What we can do at the present point is to explore *how* motion takes place.

Page 1117 * The term 'manifold' is defined in Appendix D.

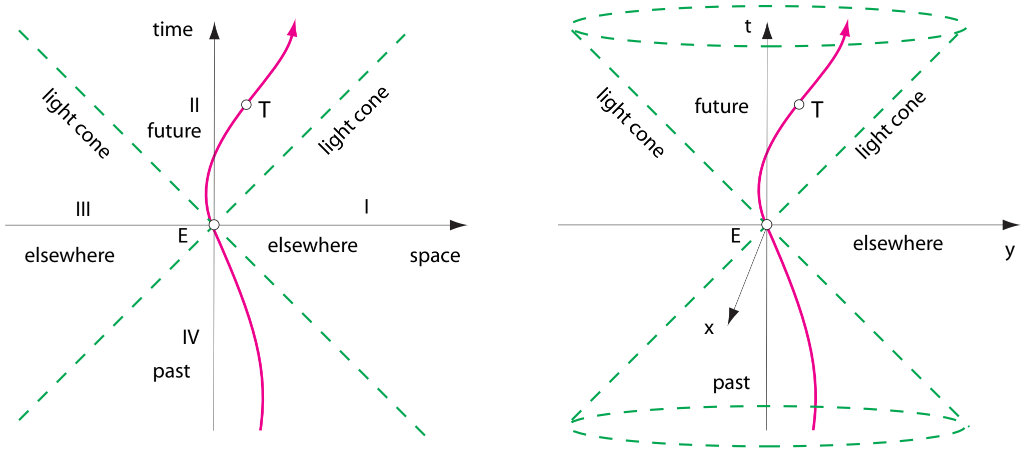


Figure 140 A space-time diagram of an object T seen from an inertial observer O in the case of one and two spatial dimensions

Can we travel to the past? – Time and causality

Given that time is different for different observers, does time nevertheless order events in sequences? The answer of relativity is a clear yes and no. Certain sets of events are not in any given sequence; others sets are. This is best seen in a space-time diagram.

Sequences of events can clearly be defined only if one event is the *cause* of another. But this connection can only apply one event exchanges energy with another event (e.g. through a signal). In other words, a relation of cause and effect between two events implies that energy or signals can travel from one event to the other; therefore, the speed connecting the two events must not be larger than the speed of light. Figure 140 shows that event E at the origin of the coordinate system can only be influenced by events in quadrant IV (the *past light cone*, when all space dimensions are included), and itself can influence only events in quadrant II (the *future light cone*). Events in quadrants I and III do not influence, nor are they influenced by event E. In other words, the light cone defines the boundary between events that *can* be ordered with respect to their origin – namely those inside the cones – and those that *cannot* – those outside the cones, happening elsewhere for all observers. (Some call all the events happening elsewhere the *present*.) In short, time orders events only *partially*. For example, for two events that are not causally connected, their simultaneity and their temporal order depends on the observer!

In particular, the past light cone gives the complete set of events that can influence what happens at the origin. One says that the origin is *causally connected* only to the past light cone. This statement reflects that any influence involves transport of energy, and thus cannot travel faster than the speed of light. Note that causal connection is an invariant concept: all observers agree on whether it applies to two given events or not. Are you able to confirm this? Challenge 563 n

A vector inside the light cone is called *timelike;* one on the light cone is called *lightlike* or *null*, and one outside the cone is called *spacelike*. For example, the *world-line* of an observer, i.e. the set of all events that make up its history, consists of timelike events only. Time is the fourth dimension; it expands space to space-time and thus 'completes' space-time. There is not much more to know about the fourth dimension, or about thinking in four dimensions.

Special relativity thus teaches us that time can be defined *only* because light cones exist. If transport of energy at speeds faster than that of light did exist, time could not be defined. Causality, i.e. the possibility of (partially) ordering events for all observers, is due to the existence of a maximal velocity.

If the speed of light could be surpassed in some way, the future could influence the past. Challenge 564 n Are you able to confirm this? In such situations one would observe *acausal* effects. However, there is an everyday experience which tells that the speed of light is indeed maximal: our memory. If the future could influence the past, we would also be able to *remember* the future. To put it in another way, if the future could influence the past, the second principle of thermodynamics would not be valid and our memory would not work.* No other data from everyday life or from experiments provide any evidence that the future can influence the past. In other words, *time travel to the past is impossible*. How the situation changes in quantum theory will be revealed later on. Interestingly, time travel to the future *is* possible, as we will see shortly.

Curiosities of special relativity

Faster than light: how far can we travel?

How far away from Earth can we travel, given that the trip should not last more than a lifetime, say 80 years, and given that we are allowed to use a rocket whose speed can approach the speed of light as closely as desired? Given the time t we are prepared to spend in a rocket, given the speed v of the rocket and assuming optimistically that it can accelerate and decelerate in a negligible amount of time, the distance d we can move away Challenge 565 ny is given by

$$d = \frac{vt}{\sqrt{1 - v^2/c^2}} . \qquad (115)$$

The distance d is larger than ct already for $v > 0.71c$, and, if v is chosen large enough, it increases beyond all bounds! In other words, relativity itself does *not* limit the distance we can travel, and not even the distance covered in a single second. We could, in principle, roam the entire universe in less than a second. In situations such as these it makes sense to introduce the concept of *proper velocity* w, defined as

$$w = d/t = \frac{v}{\sqrt{1 - v^2/c^2}} = \gamma\, v . \qquad (116)$$

* Another related result is slowly becoming common knowledge. Even if space-time had a non-trivial shape, such as a cylindrical topology with closed time-like curves, one still would not be able to travel into Ref. 242 the past, in contrast to what many science fiction novels suggest. This is made clear by Stephen Blau in a recent pedagogical paper.

As just shown, proper velocity is *not* limited by the speed of light; in fact the proper velocity of light itself is infinite.*

Synchronization and aging: can a mother stay younger than her own daughter? – Time travel to the future

A maximum speed implies that time is is different for different observers moving relative to each other. As a result we have to be careful about the way to synchronize clocks that are far apart, even if they are at rest with respect to each other in an inertial reference frame. For example, if we have two identical watches showing the same time, and if we carry one of the two for a walk and back, they will show different times afterwards. This experiment has actually been performed several times and has fully confirmed the prediction of special relativity. The time difference for a person or a watch in a plane travelling around the Earth once, at about 900 km/h, is of the order of 100 ns – not very noticeable in everyday life. In fact, the delay is easily calculated from the expression

Ref. 244, Ref. 245

$$\frac{t}{t'} = \gamma \; . \tag{118}$$

Also human bodies are clocks; they show the elapsed time, usually called *age*, by various changes in their shape, weight, hair colour, etc. If a person goes on a long and fast trip, on her return she will have aged *less* than a second person who stayed at her (inertial) home.

The most famous way to tell the story is the famous twin paradox (or clock paradox). An adventurous twin jumps on a relativistic rocket that departs from Earth for many years. Far from Earth, he jumps on another relativistic rocket coming back and returns back to Earth. The trip is illustrated in Figure 141. At his arrival, he notes that his twin brother on Earth is much older than himself. Can you explain the observation, especially the asymmetry between the two brothers? This famous result has also been confirmed in many experiments.

Ref. 249

Special relativity thus confirms, in a surprising fashion, the well-known result that those who travel a lot remain younger. The price of the retained youth is however, that everything around changes extremely quickly, in fact much faster than if one is at rest with the environment.

The twin paradox can also be seen as a confirmation of the possibility of time travel to the future. With the help of a fast rocket that comes back to its starting point, we can arrive at local times that we would never have reached within our lifetime by staying home. Alas, as has just been said, we can *never* return to the past.*

* Using proper velocity, the relation given in (108) for the superposition of two velocities $\mathbf{w}_a = \gamma_a \mathbf{v}_a$ and $\mathbf{w}_b = \gamma_b \mathbf{v}_b$ simplifies to

Challenge 566 ny

$$w_{s\|} = \gamma_a \gamma_b (v_a + v_{b\|}) \quad \text{and} \quad w_{s\perp} = w_{b\perp} \; , \tag{117}$$

where the signs $\|$ and $\perp$ designate the component in direction of motion and that perpendicular to $\mathbf{v}_a$, respectively. One can in fact write all of special relativity using 'proper' quantities, even though this is not done in this text.

Ref. 243

* There are even special books on time travel, such as the well researched text by Nahin. Note that the concept of time travel has to be clearly defined; otherwise one gets into the situation of the clerk who called his office chair a time machine, as sitting on it allows him to get to the future.

Ref. 246

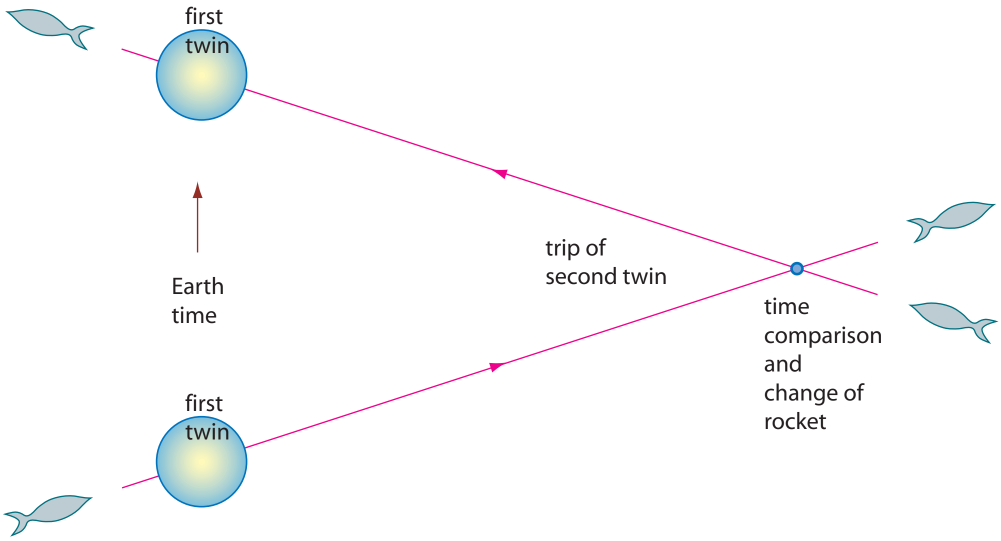


Figure 141 The twin paradox

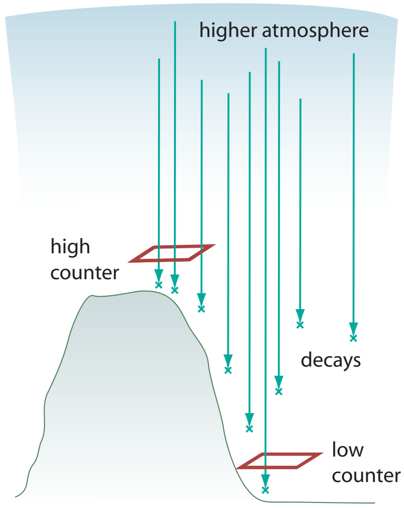


Figure 142 More muons than expected arrive at the ground because fast travel keeps young

One of the simplest experiments confirming the youth of fast travellers is the counting of muons. Muons are particles continuously formed in the up- Page 836 per atmosphere by cosmic radiation. Muons *at rest* have a finite half-life of 2.2 μs (or, at the speed of light, 660 m). After this amount of time, half of the muons are decayed. This can be measured using simple muons counters. In addition, there exist special counters that only count muons with a certain speed range, say from $0.9950c$ to $0.9954c$. One can take one of these counters on top of a mountain and put another in the valley below, as shown in Figure 142. The first time this experiment was Ref. 247 performed, the height difference was 1.9 km. Flying 1.9 km through the atmosphere at the mentioned speed takes about 6.4 μs. Using the half-life just given, this means that only about one muon in 18, Challenge 567 e or 5%, should arrive at the lower site. However, it is observed that about 72% of all muons arrive below. The reason is the relativistic time dilation. Indeed, at the mentioned speed, muons experience only a time difference of 0.71 μs during the travel from the mountain top to the valley. This shorter time yields a much lower number of lost muons than without time dilation; moreover, the measured percentage confirms the value of the predicted time dilation factor γ within experimental Challenge 568 e errors, as you may want to check. A similar effect is regularly seen when relativistic muons are produced in accelerators.

Half-life dilation has been also found for many other decaying systems, such as pions, hydrogen atoms, neon atoms and various nuclei, always confirming the predictions from special relativity. Since all bodies in nature are made of particles, the youth effect of high speeds applies to bodies of all sizes; indeed the youth effect was not only found for particles, but also for lasers, radio transmitters and clocks. Ref. 229

If motion leads to time dilation, a clock on the Equator, constantly running around the Earth, should go slower than one at the poles. However, this prediction, even though it was made by Einstein himself, is incorrect. The centrifugal acceleration leads to a reduction in gravity that exactly compensates the effect of velocity. The story reminds one to be careful when applying special relativity in situations with gravity. Special relativity is only applicable when space-time is flat, not when gravity is present. Ref. 248

In short, the question in the title of this section has a positive answer. A mother can stay younger than her daughter. We can also conclude that we cannot synchronize clocks simply by walking, clock in hand, from one place to the next. The correct way to do this is to exchange light signals. Can you describe how? Challenge 569 n

In summary, a precise definition of synchronization allows us to call two distant events simultaneous. In addition, special relativity shows that simultaneity depends on the observer. This is confirmed by all experiments performed so far.

However, the wish of the mother is not easy to realize. Let us imagine that a mother accelerates in a spaceship away from Earth with $10\,\mathrm{m/s^2}$ during a time span of ten years, then decelerates for another ten years, then accelerates for ten additional years towards the Earth, and finally decelerates for ten final years in order to land safely back on our planet. The mother took 40 years for the trip. She got as far as 22 000 light years from Earth. At her return on Earth, 44 000 years have passed. All this seems fine, until we realize that the necessary amount of fuel, even for the most efficient engine imaginable, is so large that the mass returning from the trip is only one part in $2 \cdot 10^{19}$. This amount of fuel does not exist on Earth, even if the trip is shorter. Challenge 570 e

Length contraction

The length of an object measured by an observer attached to the object is called its proper length. Special relativity makes a simple statement: the length measured by an inertial observer passing by is always smaller than the proper length. This result follows directly from the Lorentz transformations. Challenge 571 e

For a Ferrari driving at 300 km/h or 83 m/s, the length is contracted by 0.15 pm; that is less than the diameter of a proton. For the Earth, the situation is somewhat better. Seen from the Sun, the Earth moves at 30 km/s; this gives a length contraction of 6 cm. Neither of these effects has ever been measured. But larger effects could. Let us explore such situations.

Imagine a pilot flying through a barn with two doors at its two ends. The plane is slightly longer than the barn, but moves so rapidly that its relativistically contracted length is shorter than the length of the barn. Can the farmer close the barn (at least for a short time) with the plane completely inside? The answer is positive. But why can the pilot not say the following: relative to him, the barn is contracted; therefore the plane does not fit inside the barn? The answer is shown in Figure 143. For the farmer, the doors close (and reopen) at the same time. For the pilot, they do not. For the farmer, the pilot is in the dark for a short time; for the pilot, the barn is never dark. (That is not really true;

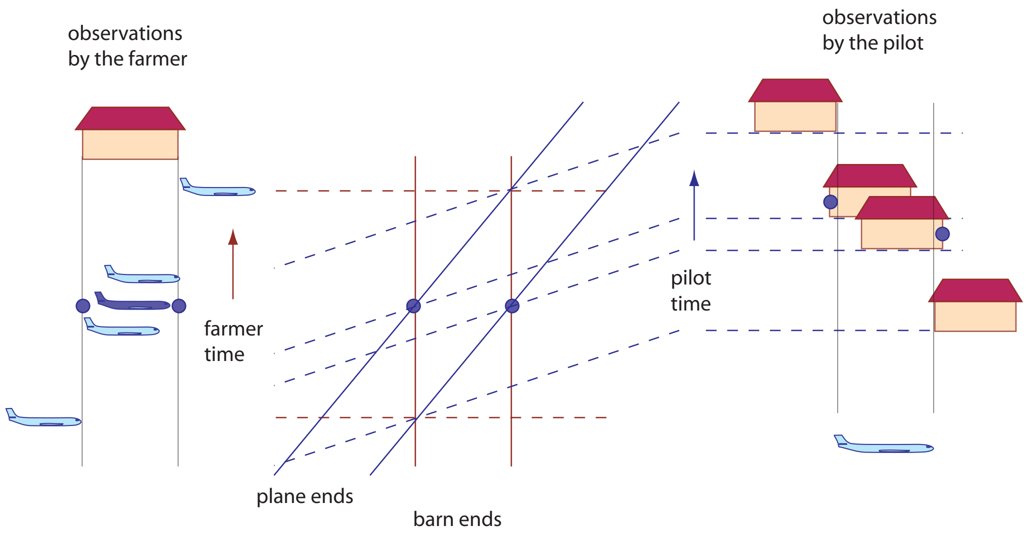


Figure 143 The observations of the rocket pilot and the barn owner

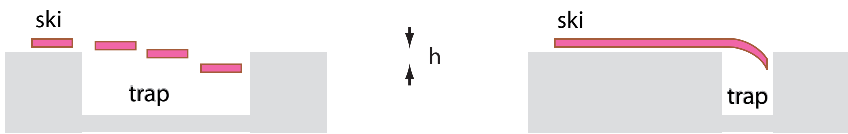


Figure 144 The observations of the trap digger and of the snowboarder, as (misleadingly) published in the literature

Challenge 572 ny can you find out the details?)

We now explore some variations of the general case. Can a rapid snowboarder fall into a hole that is a bit shorter than his board? Imagine him boarding so fast that the length contraction factor $\gamma = d/d'$ is 4.* For an observer on the ground, the snowboard is four times shorter, and when it passes over the hole, it will fall into it. However, for the boarder, it is the hole which is four times shorter; it seems that the snowboard cannot fall into it.

Ref. 250 More careful analysis shows that, in contrast to the observation of hole digger, the snowboarder does not experience the board shape as fixed; while passing over the hole, the boarder observes that the board takes on a parabolic shape and falls into the hole, as shown in Figure 144. Can you confirm this? In other words, shape is *not* an observer invariant concept. (However, rigidity is such a concept, if defined properly; can you confirm this?) Challenge 574 ny Challenge 575 ny

This story, though published, is not correct, as Harald van Lintel and Christian Gruber pointed out. One should not forget to estimate the size of the effect. At relativistic speeds the time required for the hole to affect the top of the board cannot be neglected. The result shows that the snowboarder only sees his board take on a parabolic shape if it is extremely thin and flexible. In short, at relativistic speeds the snowboarder has no time to fall any appreciable height h or to bend into the hole before passing it. Figure 144 is Ref. 251 Challenge 576 ny

Challenge 573 n * Even the Earth contracts in its direction of motion around the Sun. Is the value measurable?

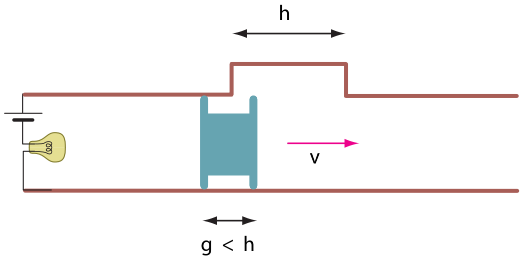


Figure 145 Does the conducting glider keep the lamp lit at large speeds?

Figure 146 What happens to the rope?

so exaggerated that it is incorrect. In practice, the snowboarder would simply speed over the hole. In other words, snowboarding does not allow to observe length contraction.

The paradoxes around *length contraction* become even more interesting in the case that one studies a conductive glider that makes electrical contact between two rails, as shown in Figure 145. The two rails are parallel, but one rail has a gap that is longer than the glider. Are you able to find out whether a lamp connected in series stays lit when the glider moves along the rails with relativistic speed? (Make the simplifying and not fully realistic assumption that electrical current flows as long and as soon as the glider touches the rails.) Do you get the same result for all observers? And what happens when the glider is longer than the detour? (Warning: this problem gives rise to *heated* debates!) What is unrealistic in this experiment?

Ref. 252

Challenge 577 n

Another example of length contraction appears when two objects, say two cars, are connected over a distance d by a straight rope. Imagine that both are at rest at time $t = 0$ and are accelerated together in exactly the same way. The observer at rest will maintain that the two cars remain the same distance apart. On the other hand, the rope needs to span a distance $d' = d/\sqrt{1 - v^2/c^2}$, which has to expand when the two cars are accelerating. In other words, the rope will break. Is this prediction confirmed by observers on each of the two cars?

Ref. 253

Challenge 578 n

A funny – but fully unrealistic – problem on length contraction is that of a submarine moving horizontally with relativistic speed. Imagine that the resting submarine has tuned its weight to float in water without any tendency to sink or to rise. Now the submarine moves (possibly with relativistic speed). The captain observes the water outside to be Lorentz contracted; thus the water is denser and he concludes that the submarine will rise. A nearby fish sees the submarine to be contracted, thus denser than water, and concludes that the submarine will sink. Who is wrong, and what is the buoyancy force? Alternatively, answer the following question that restores reality in this unrealistic problem: why is it impossible to realize a relativistic submarine?

Ref. 254

Challenge 579 n

Challenge 580 n

In summary, observing length contraction is almost always unrealistic in the case of macroscopic bodies. In contrast, it does play an important role in the case of images.

Relativistic movies – aberration and Doppler effect

We have encountered several ways in which observations change when an observer moves with high speeds. First of all, Lorentz contraction and aberration lead to *distorted* images. Second, aberration increases the viewing angle beyond the roughly 180 degrees that humans are used to in everyday life. A fast observer that looks into the direction of motion sees light rays that for a resting observer come from behind. Third, the Doppler effect produces *colour-shifted* images. Fourth, the rapid motion changes the *brightness and contrast* of the image, the so-called *searchlight effect*. Each of these changes depends on the direction of sight.

Modern computers allow to simulate the observations made by rapid observers with photographic quality, and even to produce simulated movies.* The images of Figure 147, produced by Norbert Dragon and Nicolai Mokros, allow the best understanding of the image distortion. They show the viewing angle, the circle which distinguish objects in front from those behind the observer, the coordinates of his feet and the point on the horizon toward which the observer is moving. In this way, the observations made by on observer become clear. Adding these helpful elements in your head when watching other pictures or movies helps to understand more clearly what is seen.

We note that the shape of the image seen by a moving observer is a *distorted* version of that seen by one at rest at the same point. A moving observer does not see more or different things than a resting one, in contrast to what is often suggested in cheap newspapers. Indeed, light cones are independent of observer motion.

The Lorentz contraction is measurable; however, it cannot be photographed. This distinction was discovered only in 1959. Measuring implies simultaneity at the object's position; photographing implies simultaneity at the observer's position. On a photograph, the Lorentz contraction is superposed by the effects due to different light travel times from the different parts of an object; together, they lead to a change in shape that is somewhat similar to, but not exactly the same as a rotation. The total deformation is an angle-dependent aberration; we discussed aberration at the beginning of this section. Aberration transforms circles into circles, and thus is a so-called *conformal* transformation.

Page 250

The images of Figure 149, produced by Daniel Weiskopf, also include the Doppler effect and the brightness changes. They show that these effects are at least as striking as the distortion due to aberration.

This leads to the pearl necklace paradox. If the relativistic motion transforms spheres into spheres, and rods into shorter rods, what happens to a pearl necklace moving along its own long axis? Does it get shorter or not?

Challenge 581 n

The exploration of relativistic movies is not finished yet. For example, the author predicts that interesting effects will be found simply by calculating movies of rapidly rotating spheres in motion. Also in this case, optical observation and measurement results will differ. For certain combinations of relativistic rotations and relativistic boosts, it is predicted** that the sense of rotation (clockwise or anticlockwise) will *differ* for different

Challenge 582 r

* See for example the photographic quality images and movies at http://www.anu.edu.au/Physics/Searle/ by Anthony Searle, at http://www.tat.physik.uni-tuebingen.de/~weiskopf/gallery/index.html by Daniel Weiskopf, at http://www.itp.uni-hannover.de/~dragon/stonehenge/stone1.htm by Norbert Dragon and Nicolai Mokros or at http://www.tempolimit-lichtgeschwindigkeit.de by the group of Hanns Ruder.

** In July 2005.

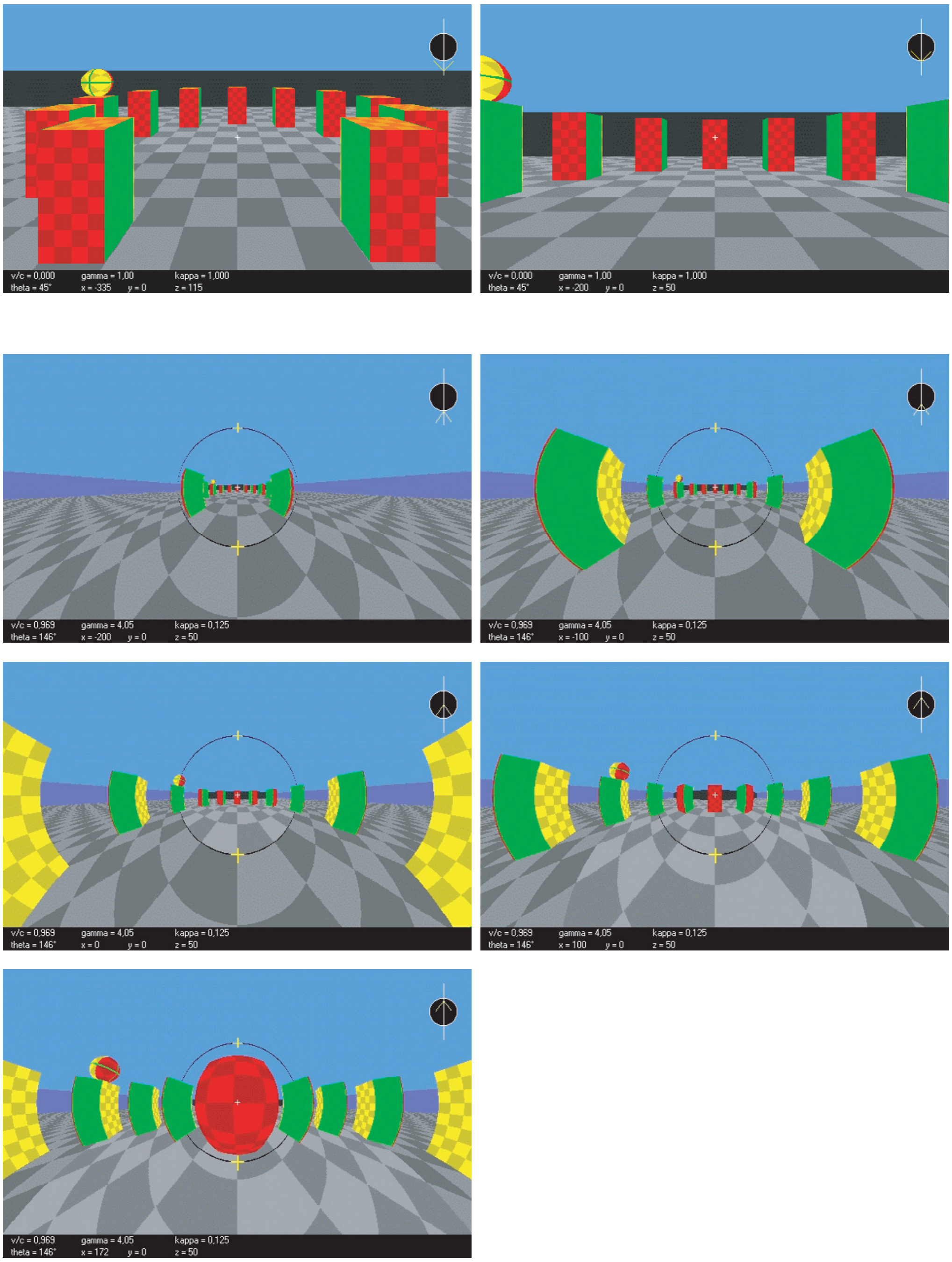


Figure 147 Flying through twelve vertical columns (shown in the two uppermost images) with 0.9 times the speed of light as visualized Nicolai Mokros and by Norbert Dragon: the effect of speed and position on distortions (courtesy of Norbert Dragon)

Figure 148 Flying through three straight and vertical columns with 0.9 times the speed of light as visualized by Daniel Weiskopf: left with the original colours, in the middle including the Doppler effect and on the right including brightness effects, thus showing what an observer would actually see

Figure 149 What a researcher standing and one running rapidly through a corridor observe (forgetting colour effects), (© Daniel Weiskopf)

observers. This effect will play an essential role in the discussion of unification.

Which is the best seat in a bus?

Ref. 253 Let us explore another surprise of special relativity. Imagine two twins inside two identically accelerated cars, starting from standstill at time $t = 0$, as described by an observer at rest with respect to both of them. Both cars contain the same amount of fuel. (There is no connecting rope now.) We easily deduce that the acceleration of the two twins stops at the same time in the frame of the outside observer, that the distance between the cars has remained the same all along for the outside observer, and that the two cars continue rolling with an identical constant velocity v, as long as friction is negligible. If we call the events at which the front car and back car engines switch off f and b, their time coordinates in the outside frame are related simply by $t_{\mathrm{f}} = t_{\mathrm{b}}$. By using the Lorentz transformations you can deduce for the frame of the freely rolling twins the relation

Challenge 583 ny

$$t_{\mathrm{b}} = \gamma \Delta x\, v/c^2 + t_{\mathrm{f}} \,, \qquad (119)$$

which means that the front twin has aged *more* than the back twin! In accelerated systems, aging is thus position dependent.

For choosing a seat in a bus, the preceding result does not help, though. It is true that the best seat in an accelerating bus is the back one, but for a decelerating bus it is the front one. At the end of a trip, the choice of a seat does not matter.

Is it correct to deduce that people on high mountains age faster than people in valleys, so that living in a valley helps avoiding grey hair?

Challenge 584 n

How fast can one walk?

To walk means to move the feet in such a way that at least one of the two feet is on the ground at any time. This is one of the rules athletes have to follow in Olympic walking competitions; they are disqualified if they break it. A student athlete was thinking about the theoretical maximum speed he could achieve in the Olympics. The ideal would be that each foot accelerates instantly to (almost) the speed of light. The highest walking speed can be achieved by taking the second foot off the ground at exactly the same instant at which the first is put down. In the beginning, by 'same instant' the student meant 'as seen by a competition judge at rest with respect to Earth'. The motion of the feet is shown in the left of Figure 150; it gives a limit speed for walking of half the speed of light. But then the student noticed that a *moving* judge will see both feet off the ground and thus disqualify the athlete for running. To avoid disqualification from *any* judge, the second foot has to wait for a light signal from the first. The limit speed for Olympic walking is thus only one third of the speed of light.

Ref. 255

Is the speed of shadow greater than the speed of light?

Contrary to what is often implied, motion faster than light does exist and is even rather common. Special relativity only constrains the motion of mass and energy. However, non-material points, non-energy transporting features and images *can* move faster than light. There are several simple examples. To be clear, we are not talking about *proper* velocity, which in these cases cannot be defined anyway. (Why?)

Page 269
Challenge 585 n

We are not talking of the situation where a particle moves faster than the velocity of light in matter, but slower than the velocity of light in vacuum. If the particle is charged, this situation gives rise to the so-called *Čerenkov radiation*. It corresponds to the v-shaped wave created by a motor boat on the sea or the cone-shaped shock wave around an aeroplane moving faster than the speed of sound. Čerenkov radiation is regularly observed;

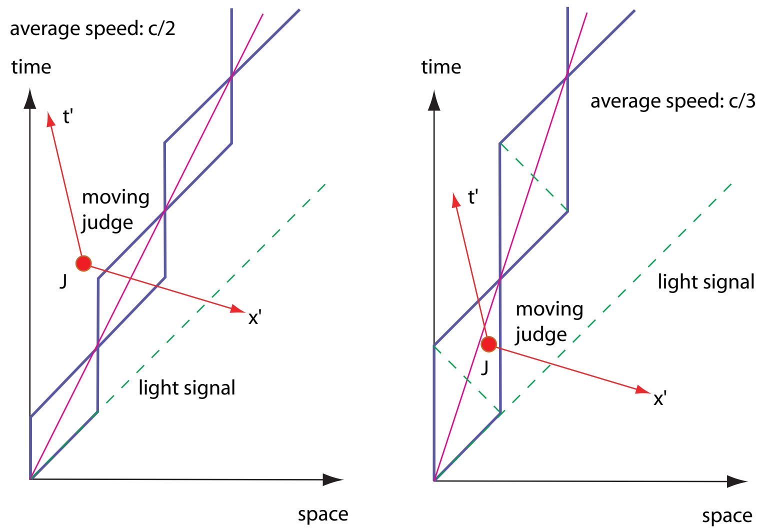


Figure 150 For the athlete on the left, the judge moving in the opposite direction sees both feet off the ground at certain times, but not for the athlete on the right

for example it is the cause of the blue glow of the water in nuclear reactors. Incidentally, the speed of light in matter can be quite low; in the centre of the Sun, the speed of light is estimated to be only around 10 km/year, and in the laboratory, for some materials, it has been found to be as low as 0.3 m/s. Ref. 256, Ref. 23

In contrast, the following examples show velocities that are genuinely faster than the velocity of light in vacuum. An example is the point marked X in Figure 151, the point at which scissors cut paper. If the scissors are closed rapidly enough, the point moves faster than light. Similar geometries can also be found in every window frame, and in fact in any device that has twisting parts.

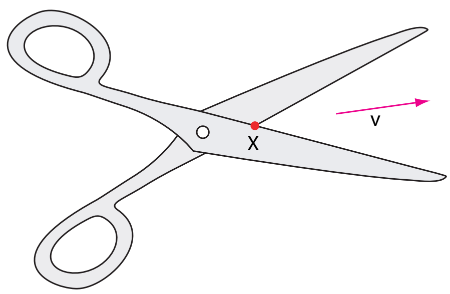


Figure 151 A simple example of motion that is faster than light

Another example of superluminal motion is the speed with which a music record – remember LPs? – disappears into its sleeve, as shown in Figure 152.

Another example appears when we remember that we live on a spherical planet. Imagine you lie on the floor and stand up. Can you show that the initial speed with which the horizon moves away from you can be larger than that of light? Challenge 586 ny

Finally, a standard example is the motion of a spot of light produced by shining a laser beam onto the Moon. If the laser is moved, the spot can easily move faster than light. The same happens for the light spot on the screen of an oscilloscope when a signal of sufficiently high frequency is fed to the input.

All these are typical examples of the *speed of shadows*, sometimes also called the *speed of darkness*. Both shadows and darkness can indeed move faster than light. In fact, there

Challenge 587 n is no limit to their speed. Can you find another example?

In addition, there is an ever-increasing number of experimental set-ups in which the phase velocity or even the group velocity of light is higher than c. They regularly make headlines in the newspapers, usually of the type 'light moves faster than light'. This surprising result is discussed in more detail later on. Page 533 Also these cases can be seen – with some imagination – as special cases for the 'speed of shadow'.

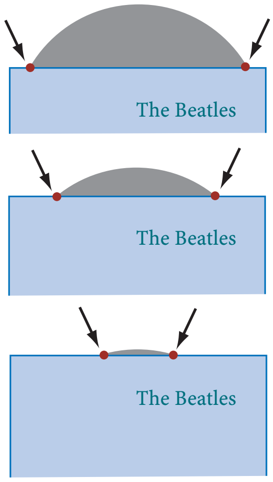


Figure 152 Another example of faster than light motion

For a different example, imagine standing at the exit of a tunnel of length l. We see a car, whose speed we know to be v, entering the other end of the tunnel and driving towards us. We know that it entered the tunnel because the car is no longer in the Sun or because its headlights were switched on at that moment. At what time t does it drive past us? Simple reasoning shows that t is given by

$$t = l/v - l/c \; . \tag{120}$$

In other words, the approaching car seems to have a velocity v_{appr} of

$$v_{\text{appr}} = \frac{l}{t} = \frac{vc}{(c-v)} \; , \tag{121}$$

which is higher than c for any car velocity v higher than $c/2$. For cars this does not happen too often, but astronomers know a type of bright object in the sky called a *quasar* (a contraction of 'quasi-stellar'), which sometimes emits high-speed gas jets. If the emission is in or near the direction to the Earth, the apparent speed – even the purely transverse component – is higher than c; such situations are now regularly observed with telescopes. Ref. 257

Note that to a second observer at the *entrance* of the tunnel, the apparent speed of the car *moving away* is given by

$$v_{\text{leav}} = \frac{vc}{(c+v)} \; , \tag{122}$$

which is *never* higher than $c/2$. In other words, objects are never seen departing with more than half the speed of light.

The story has a final twist. We have just seen that motion faster than light can be observed in several ways. But could an *object* moving faster than light be observed at all? Surprisingly, the answer is no, at least not in the common sense of the expression. First of all, since such an imaginary object, usually called a *tachyon*, moves faster than light, we can never see it approaching. If at all, tachyons can only be seen departing.

Seeing a tachyon is very similar to hearing a supersonic jet. Only *after* a tachyon has passed nearby, assuming that it is visible in daylight, could we notice it. We would first see a flash of light, corresponding to the bang of a plane passing with supersonic speed. Then we would see *two* images of the tachyon, appearing somewhere in space and departing in opposite directions, as can be deduced from Figure 153. Even if one of the two images were coming nearer, it would be getting fainter and smaller. This is, to say the least, rather

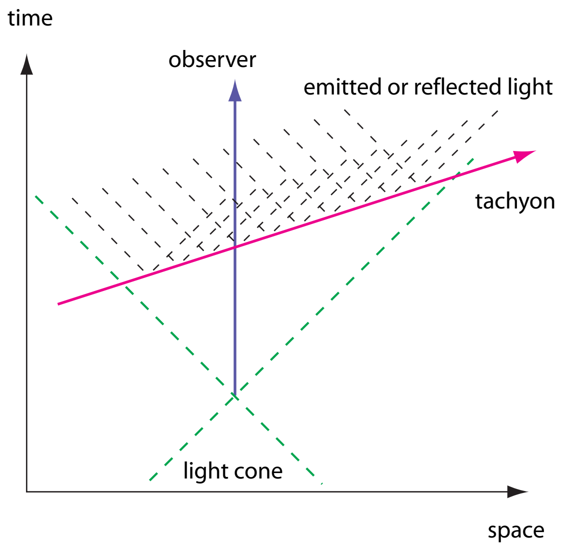


Figure 153 Hypothetical space-time diagram for tachyon observation

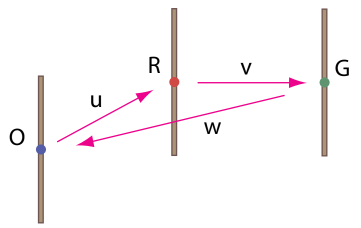


Figure 154 If O's stick is parallel to R's and R's is parallel to G's, then O's stick and G's stick are not

unusual behaviour. Moreover, if you wanted to look at a tachyon at night, illuminating it with a torch, you would have to turn your head in the direction opposite to the arm with the torch! This requirement also follows from the space-time diagram; are you able to deduce this? Nobody has ever seen such phenomena; tachyons do not exist. Tachyons would be strange objects: they would accelerate when they lose energy, and a zero-energy tachyon would be the fastest of all, with infinite speed. But no object with these properties has ever been observed. Worse, as we just saw, tachyons would seem to appear from nothing, defying laws of conservation; and note that, since tachyons cannot be seen in the usual sense, they cannot be touched either, since both processes are due to electromagnetic interactions, as we will see later in our ascent of Motion Mountain. Tachyons therefore cannot be objects in the usual sense. In the second part of our adventure we will show that quantum theory actually *rules out* the existence of (real) tachyons. However, quantum theory also requires the existence of virtual tachyons, as we will discover.

Challenge 588 e
Ref. 258
Page 288

Parallel to parallel is not parallel – Thomas rotation

Relativity has strange consequences indeed. Even though any two observers can keep a stick parallel to the stick of another, even if they are in motion with respect to each other, something strange results. A chain of sticks for which any two adjacent ones are parallel to each other will *not* ensure that the first and the last sticks are parallel. In particular, this is *never* the case if the motions of the various observers are in different directions, as is the case when the velocity vectors form a loop.

This surprising result is purely relativistic, and thus occurs *only* in the case of speeds comparable to that of light. Indeed, a general concatenation of pure boosts does not give Ref. 259 a pure boost, but a boost plus a rotation.

For example, if we walk with a stick in a fast circle, always keeping the stick parallel to the direction it had just before, at the end of the circle the stick will have an angle with respect to the original direction. Similarly, the axis of a rotating body circling a second body will *not* be pointing in the same direction after one turn, if the orbital velocity is comparable to that of light. This effect is called *Thomas precession*, after Llewellyn Thomas, who discovered it in 1925, a full 20 years after the birth of special relativity. It had escaped the attention of dozens of other famous physicists. Thomas precession is important in the inner working of atoms; we will return to it in that section of our adventure.

A never-ending story: temperature and relativity

The literature on temperature is confusing. Albert Einstein and Wolfgang Pauli agreed on the following result. The temperature T seen by an observer moving with speed v is related to the temperature T_0 measured by the observer at rest with respect to the bath via

$$T = T_0\sqrt{1 - v^2/c^2}\ . \tag{123}$$

A moving observer thus always measures lower values than a resting one. Others maintain that T and T_0 should be interchanged in this expression. Also powers other than the Ref. 260 simple square root have been proposed.

The origin of these discrepancies is simple: temperature is only defined for equilibrium situations, i.e. for baths. But a bath for one observer is not a bath for the other. For low speeds, a moving observer sees *almost* a bath; but at higher speeds the issue becomes tricky. For moving observers, there is no good way to measure temperature. The naively measured temperature value even depends on the energy range measured! In short, thermal equilibrium is not an observer-invariant concept. As a result, *no* temperature transformation formula is correct. In fact, there are not even any experimental observations that would allow the issue to be checked. Realizing such a measurement is a challenge for future experiments – but not for relativity itself.

Relativistic mechanics

As the speed of light is constant and velocities do not add up, we need to rethink the definition of mass, momentum and energy. We thus need to redo mechanics from scratch.

Mass in relativity

Page 72 In Galilean physics, the mass ratio between two bodies was defined using collisions; it was given by the negative inverse of the velocity change ratio

$$\frac{m_2}{m_1} = -\frac{\Delta v_1}{\Delta v_2} \; . \tag{124}$$

However, experiments show that the expression must be different for speeds near that of light. In fact, thinking alone can show this; are you able to do so? Challenge 589 ny

There is only one solution to this issue. The two Galilean conservation theorems $\sum_i m_i \mathbf{v}_i = \text{const}$ for momentum and $\sum_i m_i = \text{const}$ for mass have to be changed into Ref. 261

$$\sum_i \gamma_i m_i \mathbf{v}_i = \text{const} \tag{125}$$

and

$$\sum_i \gamma_i m_i = \text{const} \; . \tag{126}$$

These expressions, which are correct throughout the rest of our ascent of Motion Mountain, imply, among other things, that teleportation is *not* possible in nature. (Can you confirm this?) Obviously, in order to recover Galilean physics, the relativistic correction factors γ_i have to be equal to 1 for everyday life velocities, and have to differ noticeably from that value only for velocities near the speed of light. Even if we did not know the value of the relativistic correction, we deduce it from the collision shown in Figure 155. Challenge 590 n

In the first frame of reference we have $\gamma_v m v = \gamma_V M V$ and $\gamma_v m + m = \gamma_V M$. From the observations of the second frame of reference we deduce that V composed with V gives v, in other words, that Challenge 591 e

$$v = \frac{2V}{1 + V^2/c^2} \; . \tag{127}$$

When these equations are combined, the relativistic correction γ is found to depend on the magnitude of the velocity v through

$$\gamma_v = \frac{1}{\sqrt{1 - v^2/c^2}} \; . \tag{128}$$

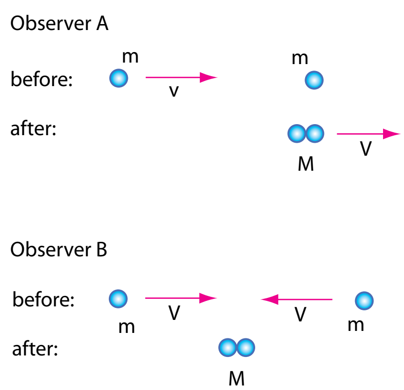


Figure 155 An inelastic collision of two identical particles seen from two different inertial frames of reference

With this expression, and a generalization of the situation of Galilean physics, the *mass* ratio between two colliding particles is defined as the ratio

$$\frac{m_1}{m_2} = -\frac{\Delta(\gamma_2 v_2)}{\Delta(\gamma_1 v_1)} \; . \tag{129}$$

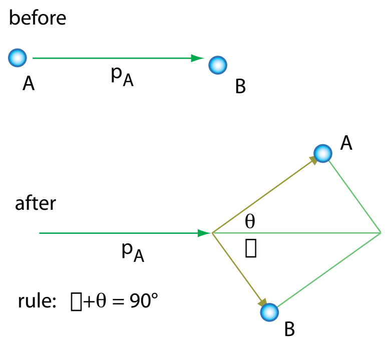


Figure 156 A useful rule for playing non-relativistic snooker

(We do not give the generalized mass definition mentioned in Galilean mechanics that is based on acceleration ratios, because it contains some subtleties that we will discover shortly.) The correction factors γ_i ensure that the mass defined by this equation is the same as the one defined in Galilean mechanics, and that it is the same for all types of collision a body may have.* In this way, the concept of mass remains a number characterizing the difficulty of accelerating a body, and it can still be used for *systems* of bodies as well.

Following the example of Galilean physics, we call the quantity

$$\mathbf{p} = \gamma m \mathbf{v} \tag{130}$$

the *(linear) relativistic (three-) momentum* of a particle. Again, the total momentum is a *conserved* quantity for any system not subjected to external influences, and this conservation is a direct consequence of the way mass is defined.

For low speeds, or $\gamma \approx 1$, the value of momentum is the same as that of Galilean physics. But for high speed, momentum increases faster than velocity, as it tends to infinity when approaching light speed. Momentum is thus not proportional to velocity at large speeds.

Why relativistic snooker is more difficult

A well-known property of collisions between a moving sphere or particle and a resting one of the *same mass* is important when playing snooker, pool or billiards. After such a collision, the two spheres will depart at a *right angle* from each other, as shown in Figure 156.

However, experiments show that the right angle rule is *not* realized for relativistic collisions. Indeed, using the conservation of momentum, you can find with a bit of dexterity

Challenge 593 ny that

$$\tan\theta \tan\varphi = \frac{2}{\gamma+1} , \tag{131}$$

where the angles are defined in Figure 157. In other words, the sum $\varphi + \theta$ is *smaller* than a right angle in the relativistic case. Relativistic speeds thus completely change the

Challenge 592 e

* The results below also show that $\gamma = 1 + T/mc^2$, where T is the kinetic energy of a particle.

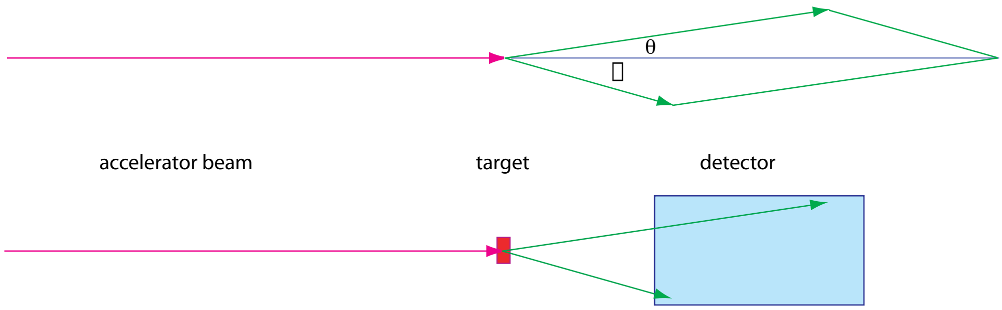


Figure 157 The dimensions of detectors in particle accelerators are based on the relativistic snooker angle rule

game of snooker. Indeed, every accelerator physicist knows this; for electrons or protons such angles can be easily deduced from photographs taken with cloud chambers, which show the tracks of particles when they fly through them. All photography confirm the above expression. In fact, the shape of detectors is chosen according to expression (131), as sketched in Figure 157. If relativity were wrong, most of these detectors would not work, as they would miss most of the particles after the collision.

Mass is concentrated energy

Let us go back to the collinear and inelastic collision of Figure 155. What is the mass M of the final system? Calculation shows that

Challenge 594 n

$$M/m = \sqrt{2(1+\gamma_v)} > 2 \,. \tag{132}$$

In other words, the mass of the final system is *larger* than the sum of the two original masses m. In contrast to Galilean mechanics, the sum of all masses in a system is *not* a conserved quantity. Only the sum $\sum_i \gamma_i m_i$ of the corrected masses is conserved.

Relativity provides the solution of this puzzle. Everything falls into place if, for the *energy* E of an object of mass m and velocity v, we use the expression

$$E = \gamma m c^2 = \frac{mc^2}{\sqrt{1 - v^2/c^2}} \,, \tag{133}$$

applying it both to the total system and to each component. The conservation of the corrected mass can then be read as the conservation of energy, simply without the factor c^2. In the example of the two identical masses sticking to each other, the two particles are thus each described by mass and energy, and the resulting system has an energy E given by the sum of the energies of the two particles. In particular, it follows that the energy E_0 of a body *at rest* and its mass m are related by

$$E_0 = mc^2 \tag{134}$$

which is perhaps the most beautiful and famous discovery of modern physics. Since the value for c^2 is so large, we can say that *mass is concentrated energy*. In other words, special relativity says that every mass has energy, and that every form of energy in a system has mass. Increasing the energy of a system increases its mass, and decreasing the energy content decreases the mass. In short, if a bomb explodes inside a closed box, the mass, weight and momentum of the box are the same before and after the explosion, but the combined mass of the debris inside the box will be *smaller* than before. All bombs – not only nuclear ones – thus take their energy from a reduction in mass. In addition, every action of a system, such a caress, a smile or a look, takes its energy from a reduction in mass.

The kinetic energy T is thus given by

$$T = \gamma mc^2 - mc^2 = \frac{1}{2}mv^2 + \frac{1 \cdot 3}{2 \cdot 4} m \frac{v^4}{c^2} + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6} \frac{v^6}{c^4} + \dots \qquad (135)$$

which reduces to the Galilean value only for low speeds.

The mass–energy equivalence $E = \gamma mc^2$ implies that *all* energy taken from matter results in a mass decrease. When a person plays the piano, thinks or runs, its mass decreases. When a cup of tea cools down or when a star shines, the mass decreases. The mass–energy equivalence pervades all of nature.

By the way, we should be careful to distinguish the transformation of *mass* into energy from the transformation of *matter* into energy. The latter is much more rare. Can you give some examples?

Challenge 595 n

The mass-energy relation (133) means the death of many science fiction fantasies. It implies that there are *no* undiscovered sources of energy on or near Earth. If such sources existed, they would be measurable through their mass. Many experiments have looked for, and are still looking for, such effects with a negative result. Free energy is unavailable in nature.*

The mass–energy relation $m = E_0/c^2$ also implies that one needs about 90 thousand million kJ (or 21 thousand million kcal) to increase one's weight by one single gram – even though diet experts have slightly different opinions on this matter. In fact, humans do get their everyday energy from the material they eat, drink and breathe by reducing its combined mass before expelling it again. However, this *chemical mass defect* appearing when fuel is burned cannot yet be measured by weighing the materials before and after the reaction; the difference is too small, because of the large conversion factor involved. Indeed, for any chemical reaction, bond energies are about 1 aJ (6 eV) per bond; this gives a weight change of the order of one part in 10^{10}, too small to be measured by weighing people or mass differences between food and excrement. Therefore, for chemical processes mass can be approximated to be constant, as is indeed done in Galilean physics and in everyday life.

Modern methods of mass measurement of *single molecules* have made it possible to measure the chemical mass defect through comparisons of the mass of a single molecule with that of its constituent atoms. David Pritchard's group has developed so-called *Penning traps* that allow masses to be determined from the measurement of frequencies;

* For example, in the universe there may still be some extremely diluted, yet undiscovered, form of energy, called *dark matter*. It is predicted from (quite difficult) mass measurements. The issue has not been finally resolved.

the attainable precision of these cyclotron resonance experiments is sufficient to confirm
Ref. 262 $\Delta E_0 = \Delta mc^2$ for chemical bonds. In future, increased precision will even allow precise bond energies to be determined in this way. Since binding energy is often radiated as light, we can say that these modern techniques make it possible to *weigh* light.

Thinking about light and its mass was also the basis for Einstein's first derivation of the mass–energy relation. When an object emits two equal light beams in opposite directions, its energy decreases by the emitted amount. Since the two light beams are equal in energy and momentum, the body does not move. If the same situation is described from the
Challenge 596 ny viewpoint of a moving observer, we get again that the *rest energy* of the object is

$$E_0 = mc^2 . \tag{136}$$

In summary, collisions and any other physical processes need relativistic treatment whenever the energy involved is a sizeable fraction of the rest energy.

How are energy and momentum related? The definitions of momentum (130) and en-
Challenge 597 e ergy (133) lead to two basic relations. First of all, their magnitudes are related by

$$m^2c^4 = E^2 - p^2c^2 \tag{137}$$

for all relativistic systems, be they objects or, as we will see below, radiation. For the momentum *vector* we get the other important relation

$$\mathbf{p} = \frac{E}{c^2}\mathbf{v} , \tag{138}$$

which is equally valid for *any* type of moving energy, be it an object or a beam or a pulse
Challenge 598 e of radiation.* We will use both relations regularly in the rest of our ascent of the Motion Mountain, including the following situation.

Collisions, virtual objects and tachyons

We have just seen that in relativistic collisions the conservation of total energy and momentum are intrinsic consequences of the definition of mass. So let us have a look at collisions in more detail, using these new concepts. Obviously a *collision* is a process, i.e. a series of events, for which

- the total momentum before the interaction and after the interaction is the same;
- the momentum is exchanged in a small region of space-time;
- for small velocities, the Galilean description is valid.

In everyday life an *impact*, i.e. a short distance interaction, is the event at which both objects change momentum. But the two colliding objects are located at *different* points
Ref. 263 when this happens. A collision is therefore described by a space-time diagram such as the one in Figure 158, reminiscent of the Orion constellation. It is easy to check that the process described by such a diagram shows all the properties of a collision.

The right-hand side of Figure 158 shows the same process seen from another, Greek, frame of reference. The Greek observer says that the first object has changed its mo-

* In 4-vector notation, we can write $v/c = \mathbf{P}/P_0$, where $P_0 = E/c$.

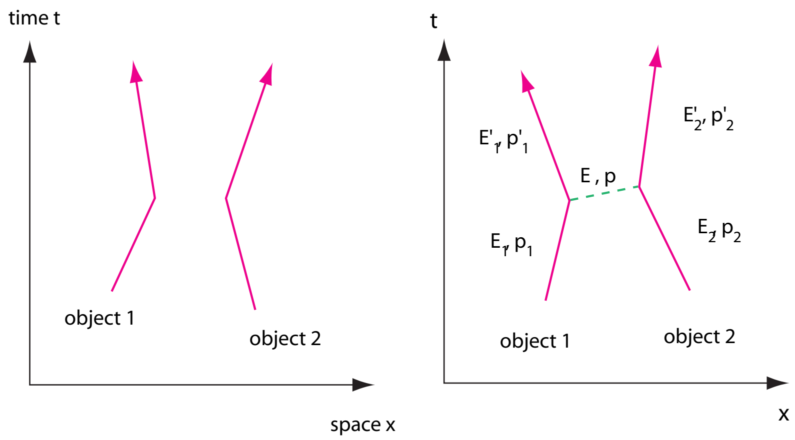


Figure 158 Space-time diagram of a collision for two observers

mentum *before* the second one. That would mean that there is a short interval when momentum and energy are *not* conserved!

The only way to save the situation is to assume that there is an exchange of a third object, drawn with a dotted line. Let us find out what the properties of this object are. If we give numerical subscripts to the masses, energies and momenta of the two bodies, and give them a prime after the collision, the unknown mass obeys

Challenge 599 ny

$$m^2c^4 = (E_1 - E_1')^2 - (p_1 - p_1')^2c^2 = 2m_1^2c^4 - 2E_1E_1'(\frac{1 - v_1v_1'}{c^2}) < 0 \,. \qquad (139)$$

This is a strange result, because a negative number means that the unknown mass is an *imaginary* number, not a real and positive one!* On top of that, we also see directly from the second graph that the exchanged object moves faster than light. It is a *tachyon*, from the Greek ταχυς 'rapid'. In other words, collisions involve motion that is faster than light! We will see later that collisions are indeed the *only* processes where tachyons play a role in nature. Since the exchanged objects appear only during collisions, never on their own, they are called *virtual* objects, to distinguish them from the usual, *real* objects, which can move freely without restriction.** We will study their properties later on, in the part of the text on quantum theory. Only virtual objects may be tachyons. Real objects are always *bradyons* – from the Greek βραδυς 'slow' – or objects moving slower than light. Note that tachyons, despite their high velocity, do not allow transport of energy faster than light,

check the greek

check the greek

* It is usual to change the mass–energy and mass–momentum relation of tachyons to $E = \pm mc^2/\sqrt{v^2/c^2 - 1}$ and $p = \pm mv/\sqrt{v^2/c^2 - 1}$; this amounts to a redefinition of m. After the redefinition, tachyons have *real* mass. The energy and momentum relations underline that (certain) tachyons lose energy and momentum when they get faster. (Provocatively, a single tachyon in a box would solve all energy problems.) Both signs for the energy and momentum relation must be retained, because otherwise the equivalence of all inertial observers would not be given. Tachyons thus do not have a minimum energy and minimum momentum: the two quantities are unbounded from below in the case of tachyons.

** More precisely, a virtual particle does not obey the relation $m^2c^4 = E^2 - p^2c^2$ valid for the real counterpart.

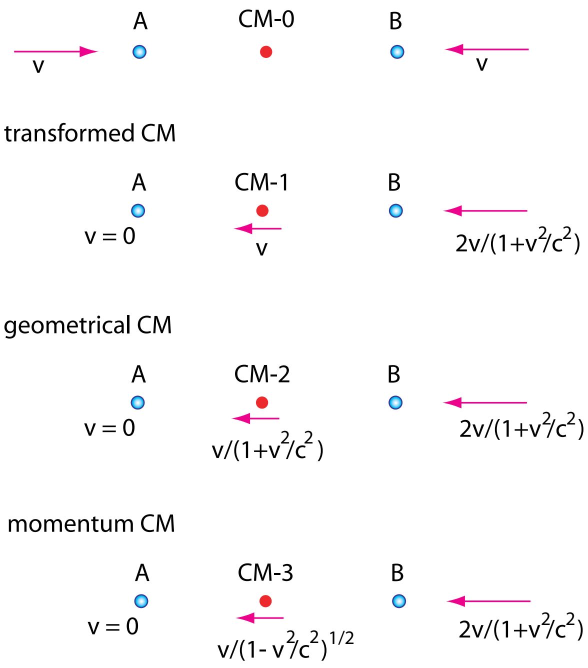


Figure 159 There is no way to define a relativistic centre of mass

and that they do not violate causality if and only if they are emitted and absorbed with the same probability. Can you confirm all this? Challenge 600 ny

There is an additional secret hidden in collisions. In the right-hand side of Figure 158, the tachyon is emitted by the first object and absorbed by the second one. However, it is easy to find an observer where the opposite happens. In short, the direction of travel of a tachyon depends on the observer! In fact, this is the first hint about *antimatter* we have encountered in our adventure. In space-time diagrams, matter and antimatter travel in opposite directions. Challenge 601 n

We will return to the topic in detail in the part of the text on quantum theory. Page 708

Studying quantum theory we will also discover that a general contact interaction between objects is not described by the exchange of a *single* virtual object, but by a continuous *stream* of virtual particles. For standard collisions of everyday objects the interaction turns out to be electromagnetic. In this case, the exchanged particles are virtual photons. In other words, when a hand touches another, when it pushes a stone, or when a mountain keeps the trees on it in place, streams of virtual photons are continuously exchanged. This is one of the strange ways in which we will need to describe nature.

Systems of particles: no centre of mass

Relativity also forces us to eliminate the cherished concept of *centre of mass*. We can see this already in the simplest example possible: that of two equal objects colliding.

Figure 159 shows that from the viewpoint in which one of two colliding particles is at rest, there are at least three different ways to define the centre of mass. In other words,

Ref. 264 the centre of mass is not an observer-invariant concept. We can deduce from the figure that the concept only makes sense for systems whose components move relative to each other with *small* velocities. For other cases, it is not uniquely definable. Will this hinder us in our ascent of the Motion Mountain? No. We are more interested in the motion of single particles than that of composite objects or systems.

Why is most motion so slow?

For most everyday cases, the time intervals measured by two different observers are practically equal; only at large relative speeds, typically at more than a few per cent of the speed of light, is a difference noted. Most such situations are microscopic. We have already mentioned the electrons inside a television tube or inside accelerators. Another example is the particles making up cosmic radiation; their high energy produced so many of the mutations that are the basis of evolution of animals and plants on this planet. Later we will discover that the particles involved in radioactivity are also relativistic.

But why don't we observe any rapid *macroscopic* bodies? Moving bodies with relativistic velocities, including observers, have a property not found in everyday life; when they are involved in a collision, part of their energy is converted into new matter via $E = \gamma mc^2$. In the history of the universe this has happened so many times that practically all the bodies still in relativistic motion are microscopic particles.

A second reason for the disappearance of rapid relative motion is radiation damping. Challenge 602 ny Can you imagine what happens to charges during collisions or to charges in a bath of light?

In short, almost all matter in the universe moves with small velocity relative to other matter. The few known counterexamples are either very old, such as the quasar jets mentioned above, or stop after a short time. The huge energies necessary for macroscopic relativistic motion are still found in supernova explosions, but they cease to exist after only a few weeks. In short the universe is mainly filled with slow motion because it is *old*. Page 415 We will determine the age shortly.

The history of the mass–energy equivalence formula by de Pretto and Einstein

Albert Einstein took several months after his first paper on special relativity to deduce the expression

$$E = \gamma mc^2 \tag{140}$$

which is often called the most famous formula of physics. He published it in a second, Ref. 221 separate paper towards the end of 1905. Arguably, the formula could have been discovered thirty years earlier from the theory of electromagnetism. Einstein was thus lucky that nobody deduced the result before him. In fact, at least one person did. In 1903 and 1904, *before* Einstein's first relativity paper, an unknown Italian engineer, Olinto De Pretto, was the first to calculate, discuss and publish the energy value $E = mc^2$.* As an engineer, De Pretto did not pursue the topic further. On the other hand, it might well be that Einstein

* Umberto Bartocci, mathematics professor of the University of Perugia in Italy, published the details of this surprising story in several papers. The full account is found in his book UMBERTO BARTOCCI, *Albert Einstein e Olinto De Pretto: la vera storia della formula più famosa del mondo*, Ultreja, Padova, 1998.

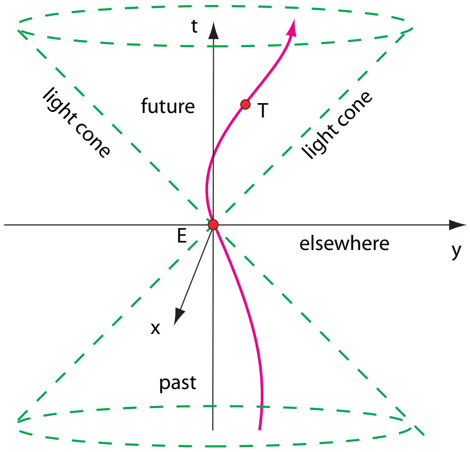


Figure 160 The space-time diagram of a moving object T

got the idea for the formula from De Pretto, possibly through his friend Michele Besso or other Italian-speaking friends he met when he visited his parents, who were living in Italy at the time. Of course, the merits of Einstein are not affected by this.

In the 1970s history repeated itself: a simple relation between the gravitational acceleration and the temperature of the vacuum was discovered, even though the result was waiting to be discovered for over 50 years. Indeed, a number of similar, anterior results were found in the libraries. Could other simple relations be hidden in modern physics?

Challenge 603 n

Four-vectors

To describe motion consistently for *all* observers, we have to introduce some new quantities. Two ideas are used. First of all, motion of particles is seen as a sequence of events. To describe events with precision, we use event coordinates, also called *4-coordinates*. These are written as

$$\mathbf{X} = (ct, \mathbf{x}) = (ct, x, y, z) = X^i . \qquad (141)$$

In this way, an event is a point in four-dimensional space-time, and is described by four coordinates. The coordinates are called the zeroth, namely time $X^0 = ct$, the first, usually called $X^1 = x$, the second, $X^2 = y$, and the third, $X^3 = z$. One can then define a *distance d* between events as the length of the difference vector. In fact, one usually uses the square of the length, to avoid writing those unwieldy square roots. In special relativity, the magnitude ('squared length') of a vector is always defined through

$$\mathbf{XX} = {X_0}^2 - {X_1}^2 - {X_2}^2 - {X_3}^2 = ct^2 - x^2 - y^2 - z^2 = X_a X^a = \eta_{ab} X^a X^b = \eta^{ab} X_a X_b . \qquad (142)$$

In this equation we have introduced for the first time two notations that are useful in relativity. First of all, we automatically sum over repeated indices. In other words, $X_a X^a$ means the sum over all products $X_a X^a$ for each index a, as just used above. Second, for every 4-vector $\mathbf{X}$ we distinguish two ways to write the coordinates, namely coordinates with superscripts and coordinates with subscripts. (In three dimensions, we only use subscripts.) They are related by the following general relation

$$X_a = \eta_{ab} X^b = (ct, -x, -y, -z) , \tag{143}$$

where we have introduced the so-called *metric* η^{ab}, an abbreviation of the matrix*

$$\eta^{ab} = \eta_{ab} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} . \tag{144}$$

Don't panic; this is all, and it won't get more difficult! We now go back to physics.

The magnitude of a position or distance vector, also called the space-time *interval*, is essentially the proper time times c^2. The *proper time* is the time shown by a clock moving in a straight line and constant velocity from the starting point to the end point in space-time. The difference from the usual 3-vectors is that the magnitude of the interval can be positive, negative or even zero. For example, if the start and end points in space-time require motion with the speed of light, the proper time is zero, as indeed is required for null vectors. If the motion is slower than the speed of light, the squared proper time is positive and the distance is timelike. For negative intervals and thus imaginary proper times, the distance is spacelike.**

Now we are ready to calculate and measure motion in four dimensions. The measurements are based on one central idea. Given the coordinates of a particle, we cannot define its velocity as the derivative of its coordinates with respect to time, since time and temporal sequences depend on the observer. The solution is to define all observables with respect to the just mentioned *proper time* τ, which is defined as the time shown by a clock attached to the object. In relativity, motion and change are always measured with respect to clocks attached to the moving system. In particular, *relativistic velocity* or *4-velocity* **U** of a body is thus defined as the change of the event coordinates or *4-coordinates* $\mathbf{X} = (ct, \mathbf{x})$ with proper time, i.e. as

$$\mathbf{U} = \mathrm{d}\mathbf{X}/\mathrm{d}\tau . \tag{145}$$

The coordinates **X** are measured in the coordinates defined by the inertial observer chosen. The value of the velocity **U** depends on the observer or coordinate system used; the velocity depends on the observer, as it does in everyday life. Using $\mathrm{d}t = \gamma\, \mathrm{d}\tau$ and thus

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma \frac{\mathrm{d}x}{\mathrm{d}t} \quad \text{, where as usual} \quad \gamma = \frac{1}{\sqrt{1 - v^2/c^2}} , \tag{146}$$

* Note that 30 % of all physics textbooks use the negative of η as metric, the so-called *spacelike convention*, and thus have negative signs in this definition. In this text, like in 70 % of all physics texts, we use the *timelike convention*.

** In the latter case, the negative of the magnitude, which then is a positive number, is called the squared *proper distance*. The proper distance is the length measured by an odometer as the object moves along that distance.

we get the relation with the 3-velocity $\mathbf{v} = \mathrm{d}\mathbf{x}/\mathrm{d}t$:

$$u^0 = \gamma c \ , \ u^{\mathrm{i}} = \gamma v_{\mathrm{i}} \quad \text{or} \quad \mathbf{U} = (\gamma c, \gamma \mathbf{v}) \ . \tag{147}$$

For small velocities we have $\gamma \approx 1$, and then the last three components of the 4-velocity are those of the usual, Galilean 3-velocity. For the magnitude of the 4-velocity $\mathbf{U}$ we find $\mathbf{UU} = U_a U^a = \eta_{ab} U^a U^b = c^2$, which is therefore independent of the magnitude of the 3-velocity $\mathbf{v}$ and makes it a timelike vector, i.e. a vector *inside* the light cone.*

Note that the magnitude of a 4-vector can be zero even though all components of such
Challenge 605 n so-called *null* vectors are different from zero. Which motions have a null velocity vector?

Similarly, the *relativistic acceleration* or *4-acceleration* $\mathbf{B}$ of a body is defined as

$$\mathbf{B} = \mathrm{d}\mathbf{U}/\mathrm{d}\tau = \mathrm{d}\mathbf{X}^2/\mathrm{d}\tau^2 \ . \tag{149}$$

Using $\mathrm{d}\gamma/\mathrm{d}\tau = \gamma \mathrm{d}\gamma/\mathrm{d}t = \gamma^4 \mathbf{va}/c^2$, we get the following relations between the four com-
Ref. 265 ponents of $\mathbf{B}$ and the 3-acceleration $\mathbf{a} = \mathrm{d}\mathbf{v}/\mathrm{d}t$:

$$B^0 = \gamma^4 \frac{\mathbf{va}}{c} \quad , \quad B^i = \gamma^2 a_i + \gamma^4 \frac{(\mathbf{va}) v_i}{c^2} \ . \tag{150}$$

The magnitude b of the 4-acceleration is rapidly found via $\mathbf{BB} = \eta_{cd} B^c B^d == -\gamma^4 (a^2 + \gamma^2 (\mathbf{va})^2/c^2) = -\gamma^6 (a^2 - (\mathbf{v} \times \mathbf{a})^2/c^2)$ and thus it does depend on the value of the 3-acceleration $\mathbf{a}$. The magnitude of the 4-acceleration is also called the *proper* acceleration because $\mathbf{B}^2 = -a^2 (v = 0)$. (What is the connection between 4-acceleration and
Challenge 606 n 3-acceleration for an observer moving with the same speed as the object?) We note that 4-acceleration lies *outside* the light cone, i.e. that it is a spacelike vector, and that $\mathbf{BU} = \eta_{cd} B^c U^d = 0$, which means that the 4-acceleration is always perpendicular to the 4-velocity.** We also note from the expression that accelerations, in contrast to velocities,

* In general, a *4-vector* is defined as a quantity (h_0, h_1, h_2, h_3), which transforms as

$$\begin{aligned} h'_0 &= \gamma_V (h_0 - h_1 V/c) \\ h'_1 &= \gamma_V (h_1 - h_0 V/c) \\ h'_2 &= h_2 \\ h'_3 &= h_3 \end{aligned} \tag{148}$$

when changing from one inertial observer to another moving with a relative velocity V in x direction; the corresponding generalization for the other coordinates are understood. This relation allows one to deduce the transformation laws for any 3-vector. Can you deduce the addition theorem (108) from this definition,
Challenge 604 n applying it to 4-velocity?

** Similarly, the *relativistic jerk* or *4-jerk* $\mathbf{J}$ of a body is defined as

$$\mathbf{J} = \mathrm{d}\mathbf{B}/\mathrm{d}\tau = \mathrm{d}^2\mathbf{U}/\mathrm{d}\tau^2 \ . \tag{151}$$

Challenge 607 ny For the relation with the 3-jerk $\mathbf{j} = \mathrm{d}\mathbf{a}/\mathrm{d}t$ we then get

$$\mathbf{J} = (J^o, J^i) = \left(\frac{\gamma^5}{c} (\mathbf{jv} + a^2 + 4\gamma^2 \frac{(\mathbf{va})^2}{c^2}) \ , \gamma^3 j_i + \frac{\gamma^5}{c^2} ((\mathbf{jv}) v_i + a^2 v_i + 4\gamma^2 \frac{(\mathbf{va})^2 v_i}{c^2} + 3(\mathbf{va}) a_i) \right) \tag{152}$$

Challenge 608 ny which we will use later on. Surprisingly, $\mathbf{J}$ does not vanish when $\mathbf{j}$ vanishes. Why not?

cannot be called relativistic; the difference between b_i and a_i or between their two magnitudes does not depend on the value of a_i, but only on the value of the speed v. In other words, accelerations require relativistic treatment only when the involved velocities are relativistic. If the velocities involved are low, even the highest accelerations can be treated with Galilean methods.

Page 301 We note that when the acceleration $\mathbf{a}$ is parallel to the speed $\mathbf{v}$, we get $B = \gamma^3 a$; when $\mathbf{a}$ is perpendicular to $\mathbf{v}$, as in circular motion, we get $B = \gamma^2 a$. We use this below.

Four-momentum

To describe motion, we also need the concept of momentum. The *4-momentum* is defined by setting

$$\mathbf{P} = m\mathbf{U} \tag{153}$$

and is therefore related to 3-momentum $\mathbf{p}$ by

$$\mathbf{P} = (\gamma mc, \gamma m\mathbf{v}) = (E/c, \mathbf{p}) . \tag{154}$$

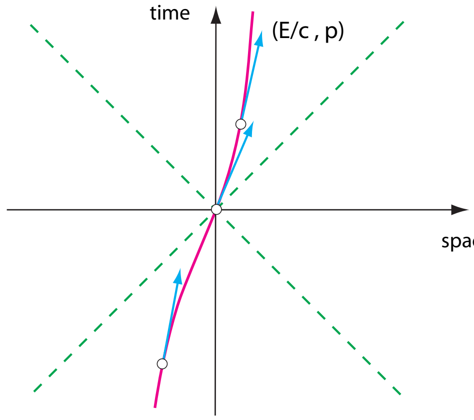


Figure 161 Energy–momentum is tangent to the world line

For this reason 4-momentum is also called the *energy–momentum 4-vector*. In short, *the 4-momentum of a body is given by mass times 4-displacement per proper time*. This is the simplest possible definition of momentum and energy. The energy–momentum 4-vector, also called *momenergy*, like the 4-velocity, is *tangent* to the world line of a particle. This follows directly from the definition, since

$$(E/c, \mathbf{p}) = (\gamma mc, \gamma m\mathbf{v}) = m(\gamma c, \gamma \mathbf{v}) = m(\mathrm{d}t/\mathrm{d}\tau, \mathrm{d}\mathbf{x}/\mathrm{d}\tau) . \tag{155}$$

The (square of the) length of momenergy, namely $\mathbf{PP} = \eta_{ab}P^a P^b$, is by definition the same for all inertial observers and found to be

$$E^2/c^2 - p^2 = m^2c^2 , \tag{156}$$

thus confirming a result given above. We have already mentioned that energies or situations are called *relativistic* if the kinetic energy $T = E - E_o$ is not negligible when compared to the rest energy $E_0 = mc^2$. A particle whose kinetic energy is much higher than its rest mass is called *ultrarelativistic*. Particles in accelerators or in cosmic rays fall into Challenge 609 n this category. (What is their energy–momentum relation?)

In contrast to Galilean mechanics, relativity specifies an absolute zero for the energy. One cannot extract more energy than mc^2 from a system of mass m. In particular, a zero value for potential energy is fixed in this way. In short, relativity shows that energy is bounded from below.

Note that by the term 'mass' m we always mean what is sometimes also called the *rest mass*. This name derives from the bad habit of many science fiction and high-school books of calling the product γm the *relativistic mass*. Workers in the field usually (but not unanimously) reject this concept, as did Einstein himself, and they also reject the often heard sentence that '(relativistic) mass increases with velocity'. Relativistic mass and energy would then be two words for the same concept. This last statement is at the level of the tabloid press, and not worthy of any motion expert.

Ref. 266

Not all Galilean energy contributes to mass. Potential energy in an outside field does not count. Relativity forces us to precise energy booking. 'Potential energy' is an abbreviation for 'energy reduction of the outside field'.

Can you show that for two particles with momenta P_1 and P_2, one has $P_1 P_2 = m_1 E_2 = M_2 E_1 = c^2 \gamma v_{12} m_1 m_2$, where v_{12} is their relative velocity?

Challenge 610 n

Four-force

We note that 4-force $\mathbf{K}$ is defined as

$$\mathbf{K} = \mathrm{d}\mathbf{P}/\mathrm{d}\tau = m\mathbf{B} \tag{157}$$

and that, therefore, contrary to an often heard statement, force remains mass times acceleration in relativity. From the definition of $\mathbf{K}$ we deduce the relation with 3-force $\mathbf{f} = \mathrm{d}\mathbf{p}/\mathrm{d}t = m\mathrm{d}(\gamma\mathbf{v})/\mathrm{d}t$, namely*

Ref. 265, Ref. 267

$$\mathbf{K} = (K^{\mathrm{o}}, K^{\mathrm{i}}) = (\gamma^4 m\mathbf{va}/c, \gamma^2 m a_{\mathrm{i}} + \gamma^4 v_{\mathrm{i}} \frac{m\mathbf{va}}{c^2}) = (\frac{\gamma}{c}\frac{\mathrm{d}E}{\mathrm{d}t}, \gamma\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t}) = (\gamma\frac{\mathbf{fv}}{c}, \gamma\mathbf{f}) \; . \tag{158}$$

Challenge 612 ny

Also the 4-force, like the 4-acceleration, is orthogonal to the 4-velocity. The meaning of the zeroth component of the 4-force can be easily recognized: it is the *power* required to accelerate the object. One has $\mathbf{KU} = c^2\mathrm{d}m/\mathrm{d}\tau = \gamma^2(\mathrm{d}E/\mathrm{d}t - \mathbf{fv})$; this is the proper rate at which the internal energy of a system increases. The product $\mathbf{KU}$ vanishes only for rest-mass conserving forces. Particle collisions that lead to reactions do not belong to this class. In everyday life, the rest mass is preserved, and then one gets the Galilean expression $\mathbf{fv} = \mathrm{d}E/\mathrm{d}t$. We now turn to a different topic.

Rotation in relativity

If at night we turn around our own axis while looking at the sky, the stars move with a much higher velocity than that of light. Most stars are masses, not images. Their speed should be limited by that of light. How does this fit with special relativity?

The example helps to clarify in another way what the limit velocity actually is. Physically speaking, a rotating sky does *not* allow superluminal energy transport, and thus is not in contrast with the concept of a limit speed. Mathematically speaking, the speed of light limits relative velocities *only* between objects that come *near* to each other. To compare

* Some authors define 3-force as $\mathrm{d}\mathbf{p}/\mathrm{d}\tau$; then $\mathbf{K}$ looks slightly different. In any case, it is important to note that the in relativity, 3-force $\mathbf{f} = \mathrm{d}\mathbf{p}/\mathrm{d}t$ is indeed proportional to 3-acceleration $\mathbf{a}$; however, force and acceleration are not parallel to each other. In fact, for rest-mass preserving forces one finds $\mathbf{f} = \gamma m\mathbf{a} + (\mathbf{fv})\mathbf{v}/c^2$. In contrast, in relativity 3-momentum is not proportional to 3-velocity, but parallel to it.

Challenge 611 n

velocities of distant objects is only possible if all velocities involved are constant in time; this is not the case in the present example. Avoiding this limitation is one of the reasons to prefer the differential version of the Lorentz transformations. In many general cases Page 280 relative velocities of *distant* objects can be higher than the speed of light. We encountered a first example above, when discussing the car in the tunnel, and we will encounter a few Page 308 additional examples shortly.

With this clarification, we can now have a short look at *rotation* in relativity. The first question is how lengths and times change in a rotating frame of reference. You may want to check that an observer in a rotating frame agrees with a non-rotating colleague on the radius of a rotating body; however, both find that the rotating body, even if it is rigid, has a *different circumference* from before it started rotating. Sloppily speaking, the value of π *changes* for rotating observers. The ratio between the circumference c and the radius r turns out to be $c/r = \gamma 2\pi$; it increases with rotation speed. This counter-intuitive result is often called *Ehrenfest's paradox*. Among others, it shows that space-time for an observer on a rotating disc is *not* the Minkowski space of special relativity.

Challenge 613 ny

Challenge 614 ny

Ref. 268

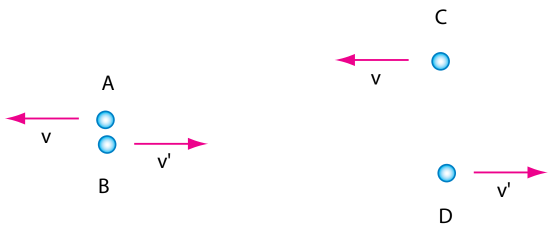


Figure 162 On the definition of relative velocity

Rotating bodies behave strangely in many ways. For example, one gets into trouble when one tries to synchronize clocks mounted on a circle around the rotation centre. If one starts synchronizing the clock at O_2 with that at O_1, continuing up to clock O_n, one finds that the last clock is *not* synchronized with the first. This result reflects the change in circumference just mentioned. In fact, a careful study shows that the measurements of length and time intervals lead all observers O_k to conclude that they live in a rotating space-time. Rotating disks can thus be used as an introduction to general relativity, where this curvature and its effects form the central topic. More about this in the next chapter.

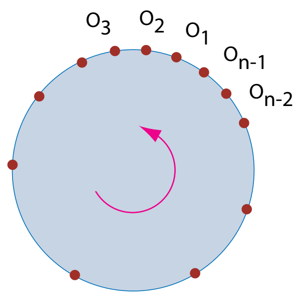


Figure 163 Observers on a rotating object

Is angular velocity limited? Yes; the tangential speed in an inertial frame of reference cannot exceed that of light. The limit thus depends on the *size* of the body in question. That leads to a neat puzzle: can one *see* objects rotating very rapidly?

Challenge 615 ny

We mention that 4-angular momentum is defined naturally as

$$l^{ab} = x^a p^b - x^b p^a \; . \tag{159}$$

In other words, 4-angular momentum is a *tensor*, not a vector, as shown by its two indices. Angular momentum is also obviously conserved in special relativity, so that there are no surprises on this topic. As usual, the moment of inertia is defined as the proportionality factor between angular velocity and angular momentum.

Challenge 616 ny

Obviously, for a rotating particle, the rotational energy is part of the rest mass. You

Challenge 617 ny may want to calculate the fraction for the Earth and the Sun. It is not large. By the way,
Challenge 618 ny how would you determine whether a small particle, too small to be seen, is rotating?

In relativity, rotation and translation combine in strange ways. Imagine a cylinder in uniform rotation along its axis, as seen by an observer at rest. As Max von Laue discussed, the cylinder will appear *twisted* to an observer moving along the rotation axis. Can you
Challenge 619 ny confirm this?

Wave motion

Waves in Galilean physics are described by wave vector and frequency. In special relativity, the two are combined in the wave 4-vector given by

$$\mathbf{L} = \frac{1}{\lambda}\left(\frac{\omega}{c}, \mathbf{n}\right) \qquad (160)$$

where λ is the wavelength, ω the wave velocity, and $\mathbf{n}$ the normed direction vector. An observer with 4-velocity $\mathbf{U}$ finds that a wave $\mathbf{L}$ has frequency ν. Can you show that

$$\nu = \mathbf{LU} \qquad (161)$$

Ref. 226 Interestingly, the wave velocity ω transforms in a different way than particle velocity except in the case $\omega = c$. Also the aberration formula for wave motion differs from that for
Challenge 620 ny particles, except in the case $\omega = c$.

The action of a free particle – how do things move?

If we want to describe relativistic motion of a free particle with an extremal principle,
Page 161 we need a definition of the action. We already know that physical action measures the change occurring in a system. For an inertially moving or free particle, the only change is the ticking of its proper clock. As a result, the action of a free particle will be proportional to the elapsed proper time. In order to get the standard unit of energy times time, or Js, for the action, the first guess for the action of a free particle is

$$S = -mc^2 \int_{\tau_1}^{\tau_2} d\tau \,, \qquad (162)$$

where τ is the proper time along its path. This is indeed the correct expression; energy and momentum conservation follow from it, as the proper time is maximal for straight-
Challenge 621 ny line motion with constant velocity. Can you confirm this? Indeed, in nature, all particles move in such a way that their proper time is maximal. In other words, we again find that in nature things change as little as possible. Nature is like a wise old man: its motions are as slow as possible. If you prefer, every change is maximally effective. As we mentioned already before, Bertrand Russell called this the *law of cosmic laziness*.

The action can also be written in more complex ways, in order to frighten the hell out of readers. These other, equivalent ways to write it prepare for the future, in particular for general relativity:

$$S = \int L \, \mathrm{d}t = -mc^2 \int_{t_1}^{t_2} \frac{1}{\gamma} \, \mathrm{d}t = -mc \int_{\tau_1}^{\tau_2} \sqrt{u_a u^a} \, \mathrm{d}\tau = -mc \int_{s_1}^{s_2} \sqrt{\eta^{ab} \frac{\mathrm{d}x_a}{\mathrm{d}s} \frac{\mathrm{d}x_b}{\mathrm{d}s}} \, \mathrm{d}s \,, \tag{163}$$

where s is some arbitrary, but monotonically increasing function of τ – such as τ itself – and the *metric* $\eta^{\alpha\beta}$ of special relativity is given as usual as

$$\eta^{ab} = \eta_{ab} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} . \tag{164}$$

You can easily confirm the form of the action by deducing the equation of motion with the usual procedure. Challenge 622 ny

In short, nature is in not a hurry: every object moves in a such way that its own clock shows the *longest* delay possible, compared with any alternative motion nearby.* This general principle is also valid for particles under the influence of gravity, as we will see in the section on general relativity, and under the influence of electric or magnetic interactions. In fact, it is valid in all (macroscopic) cases of motion found in nature. In nature, *proper time is always maximal.* Alternatively, things move along paths of *maximal aging*.

Later in our walk, we will extend the action to include interactions. Here we just note that the longest proper time is realized when the difference between kinetic and potential energy is minimal. Can you confirm this? Challenge 624 ny For the case of Galilean physics the longest proper time thus indeed implies the smallest average difference between the two energy types. This is the principle of least action in its Galilean formulation.

Page 161 Earlier on, we saw that the action measures the change going on in a system. Minimizing proper time is the way that nature minimizes change. We thus again find that nature is the opposite of a Hollywood movie; nature changes in the most economical way possible. Speculating on the deeper meaning of this result is left to your personal preferences; enjoy it!

Conformal transformations: Why is the speed of light constant?

The distinction between space and time in special relativity depends on the observer. On the other hand, all inertial observers do agree on the position, shape and orientation of the light cone at a point. The light cones at each point thus are the basic physical 'objects' with which space-time is described in the theory of relativity. Given the importance of light cones, we might ask if inertial observers are the only ones that observe the same light cones. Interestingly, it turns out that there are *other* such observers.

The first group of these additional observers is made up of those using different units of measurement, namely units in which all time and length intervals are multiplied by a

* If neutrinos were massless, the action would not work for them. Why? Can you find an alternative for this (admittedly academic) case? Challenge 623 ny

scale factor λ. The transformations among these points of view are given by

$$x_a \mapsto \lambda x_a \tag{165}$$

and are called *dilations*.

A second type of additional observers are found by applying the so-called *special conformal transformations*. They are combinations of an *inversion*

$$x_a \mapsto \frac{x_a}{x^2} \tag{166}$$

with a *translation* by a vector b_a, namely

$$x_a \mapsto x_a + b_a \ , \tag{167}$$

and a second inversion. This gives for the expression for the special conformal transformations

$$x_a \mapsto \frac{x_a + b_a x^2}{1 + 2b_a x^a + b^2 x^2} \quad \text{or} \quad \frac{x_a}{x^2} \mapsto \frac{x_a}{x^2} + b_a \ . \tag{168}$$

These transformations are called *conformal* because they do not change angles of (infin-
Challenge 625 ny itesimally) small shapes, as you may want to check. The transformations thus leave the *form* (of infinitesimally small objects) unchanged. For example, they transform infinitesimal circles into infinitesimal circles. They are called *special* because the *full* conformal group includes the dilations and the inhomogeneous Lorentz transformations as well.*

The way in which special conformal transformations leave light cones invariant is rather subtle.

– CS – Text to be filled in. – CS –

Note that, since dilations do not commute with time translations, there is no conserved quantity associated with this symmetry. (The same happens with Lorentz boosts; in contrast, rotations and spatial translations do commute with time translations and thus do lead to conserved quantities.)

In summary, vacuum is conformally invariant – in the special way just mentioned – and thus also dilation invariant. This is another way to say that vacuum alone is not sufficient to define lengths, as it does not fix a scale factor. As expected, matter is necessary to do so. Indeed, (special) conformal transformations are not symmetries of situations containing matter. Only vacuum is conformally invariant; nature as a whole is not.

However, conformal invariance, or the invariance of light cones, is sufficient to allow velocity measurements. Obviously, conformal invariance is also *necessary* for velocity

Challenge 626 ny * The set of all *special* conformal transformations forms a group with four parameters; adding dilations and the inhomogeneous Lorentz transformations one gets fifteen parameters for the *full* conformal group. The conformal group is locally isomorphic to SU(2,2) and to the simple group SO(4,2); these concepts are
Page 1098 explained in Appendix D. Note that all this is true only for *four* space-time dimensions; in *two* dimensions, the other important case, especially in string theory, the conformal group is isomorphic to the group of arbitrary analytic coordinate transformations, and is (thus) infinite-dimensional.

Challenge 627 ny measurements, as you might want to check.

We saw that conformal invariance includes inversion symmetry. Inversion symmetry means that the large and small scales of a vacuum are related. This suggest that the constancy of the speed of light is related to the existence of inversion symmetry. This mysterious connection gives us a glimpse of the adventures we will encounter in the third part of our ascent of Motion Mountain. Conformal invariance turns out to be an important property that will lead to incredible surprises.*

Accelerating observers

So far, we have only studied what inertial, or free-flying, observers say to each other when they talk about the same observation. For example, we saw that moving clocks always run slow. The story gets even more interesting when one or both of the observers are accelerating.

One sometimes hears that special relativity cannot be used to describe accelerating observers. That is wrong: the argument would imply that even Galilean physics could not be used for accelerating observers, in contrast to everyday experience. Special relativity's only limitation is that it cannot be used in non-flat, i.e. curved, space-time. Accelerating bodies do exist in flat space-times, and therefore they can be discussed in special relativity.

As an appetizer, let us see what an accelerating, Greek, observer says about the clock of an inertial, Roman, one, and vice versa. Ref. 269 Assume that the Greek observer moves along $\mathbf{x}(t)$, as observed by the inertial Roman one. In general, the Roman/Greek clock rate ratio is given by $\Delta\tau/\Delta t = (\tau_2 - \tau_1)/(t_2 - t_1)$, where the Greek coordinates are constructed with a simple procedure: take the set of events defined by $t = t_1$ and $t = t_2$, and determine where these sets intersect the time axis of the Greek observer, and call them τ_1 and τ_2.** We assume that the Greek observer is inertial and moving with velocity v as observed by the Roman one. The clock ratio of a Greek observer is then given by

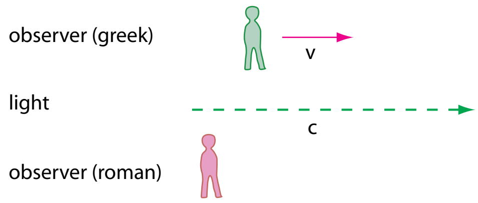


Figure 164 The simplest situation for an inertial and an accelerated observer

$$\frac{\Delta\tau}{\Delta t} = \frac{d\tau}{dt} = \sqrt{1 - v^2/c^2} = \frac{1}{\gamma_v} , \qquad (169)$$

* The conformal group does not appear only in the kinematics of special relativity; it is the symmetry group of all physical interactions, such as electromagnetism, provided that all the particles involved have zero mass, as is the case for the photon. Any field that has mass cannot be conformally invariant; therefore conformal invariance is not an exact symmetry of all of nature. Can you confirm that a mass term $m\varphi^2$ in a Lagrangian is not conformally invariant? Challenge 628 ny

However, since all particles observed up to now have masses that are many orders of magnitude smaller than the Planck mass, from a global viewpoint it can be said that they have almost vanishing mass; conformal symmetry then can be seen as an *approximate* symmetry of nature. In this view, all massive particles should be seen as small corrections, or perturbations, of massless, i.e. conformally invariant, fields. Therefore, for the construction of a fundamental theory, conformally invariant Lagrangians are often assumed to provide a good starting approximation.

** These sets form what mathematicians call *hypersurfaces*.

Challenge 629 ny as we are now used to. We find again that moving clocks run slow.

Ref. 269 For accelerated motions, the differential version of the reasoning is necessary. In other words, the Roman/Greek clock rate ratio is again $\mathrm{d}\tau/\mathrm{d}t$, and τ and $\tau + \mathrm{d}\tau$ are calculated in the same way as just defined from the times t and $t + \mathrm{d}t$. Assume again that the Greek observer moves along $\mathbf{x}(t)$, as measured by the Roman one. We find directly that

$$\tau = t - \mathbf{x}(t)\mathbf{v}(t)/c^2 \tag{170}$$

and thus

$$\tau + \mathrm{d}\tau = (t + \mathrm{d}t) - [\mathbf{x}(t) - \mathrm{d}t\mathbf{v}(t)][\mathbf{v}(t) + \mathrm{d}t\mathbf{a}(t)]/c^2 \ . \tag{171}$$

Together, this yields

$$\text{‘}\mathrm{d}\tau/\mathrm{d}t\text{’} = \gamma_v(1 - \mathbf{v}\mathbf{v}/c^2 - \mathbf{x}\mathbf{a}/c^2) \ , \tag{172}$$

a result showing that accelerated clocks can run *fast* instead of slow, depending on their position $\mathbf{x}$ and the sign of their acceleration $\mathbf{a}$. There are quotes in the expression because we see directly that the Greek observer notes

$$\text{‘}\mathrm{d}t/\mathrm{d}\tau\text{’} = \gamma_v \ , \tag{173}$$

which is *not* the inverse of equation (172). This difference becomes most apparent in the simple case of two clocks with the same velocity, one of which is accelerated constantly towards the origin with magnitude g, whereas the other moves inertially. We then have

$$\text{‘}\mathrm{d}\tau/\mathrm{d}t\text{’} = 1 + gx/c^2 \tag{174}$$

and

$$\text{‘}\mathrm{d}t/\mathrm{d}\tau\text{’} = 1 \ . \tag{175}$$

We will encounter this situation shortly. But first we clarify the concept of acceleration.

Acceleration for inertial observers

Accelerations behave differently from velocities under change of viewpoint. Let us first take the simple case in which everything moves along the x-axis: the object and two inertial observers. If a Roman inertial observer measures an acceleration $a = \mathrm{d}v/\mathrm{d}t = \mathrm{d}^2x/\mathrm{d}t^2$, and the Greek observer, also *inertial* in this case, an acceleration $\alpha = \mathrm{d}\omega/\mathrm{d}\tau =$

Ref. 270 $\mathrm{d}^2\xi/\mathrm{d}\tau^2$, we get

$$\gamma_v^3 a = \gamma_\omega^3 \alpha \ . \tag{176}$$

The relation shows that accelerations are *not* Lorentz invariant; they are so only if the velocities are small compared to the speed of light. This is in contrast to our everyday experience, where accelerations are independent of the speed of the observer.

Expression (176) simplifies in the case that the accelerations are measured at a time t in which ω vanishes – i.e. measured by the so-called *comoving* inertial observer. In that case the acceleration relation is given by

$$a_c = a\gamma_v^3 \tag{177}$$

and the acceleration $a_c = \alpha$ is also called proper acceleration, as its value describes what the Greek, comoving observer *feels*; proper acceleration describes the experience of being pushed into the accelerating seat.

In general, the observer speed and the acceleration are not collinear. One deduces how the value of 3-acceleration **a** measured by a general inertial observer is related to the value $\mathbf{a}_c$ measured by the comoving observer using expressions (150) and (148). One gets the generalization of (177)

Ref. 271

$$\mathbf{v}\mathbf{a}_c = \mathbf{v}\mathbf{a}\gamma_v^3 \tag{178}$$

and

$$\mathbf{a} = \frac{1}{\gamma_v^2}\left(\mathbf{a}_c - \frac{(1-\gamma_v)(\mathbf{v}\mathbf{a}_c)\mathbf{v}}{v^2} - \frac{\gamma_v(\mathbf{v}\mathbf{a}_c)\mathbf{v}}{c^2}\right) . \tag{179}$$

Squaring yields the relation

$$a^2 = \frac{1}{\gamma_v^4}\left(a_c^2 - \frac{(\mathbf{a}_c\mathbf{v})^2}{c^2}\right) \tag{180}$$

Page 293

which we know already in similar form. It shows (again) that the comoving or proper 3-acceleration is always larger than the 3-acceleration measured by an outside inertial observer. The faster the outside inertial observer is, the smaller the acceleration he observes. Acceleration is indeed not a relativistic invariant. The expression also shows that whenever the speed is perpendicular to the acceleration, a boost yields a factor γ_v^2, whereas a speed collinear with the acceleration gives the already mentioned γ_v^3 dependence.

Challenge 630 e

In summary, acceleration complicates many issues and requires a deeper investigation. To keep matters simple, from now on we only study *constant* accelerations. Interestingly, this situation is also a good introduction to black holes and, as we will see shortly, to the universe as a whole.

Page 441

Accelerating frames of reference

How do we check whether we live in an inertial frame of reference? An *inertial frame (of reference)* has two properties: first, the speed of light is constant. In other words, for any two observers in that frame the ratio c between twice the distance measured with a ruler and the time taken by light to travel from one point to another and back again is always the same. The ratio is independent of time and of the position of the observers. Second, lengths and distances measured with a ruler are described by Euclidean geometry. In other words, rulers behave as in daily life; in particular, distances found by counting how many rulers (rods) have to be laid down end to end, the so-called *rod distances*, behave as in everyday life. For example, they follow Pythagoras' theorem in the case of right-angled triangles.

Equivalently, an inertial frame is one for which all clocks always remain synchronized and whose geometry is Euclidean. In particular, in an inertial frame all observers at fixed coordinates always remain at rest with respect to each other. This last condition is, however, a more general one. Interestingly, there are other, non-inertial, situations where this is the case.

Non-inertial frames, or *accelerating frames*, are useful concepts special relativity. In fact, we all live in such a frame. We can use special relativity to describe it in the same way that we used Galilean physics to describe it at the beginning of our journey.

A general *frame of reference* is a continuous set of observers remaining at rest with respect to each other. Here, 'at rest with respect to each other' means that the time for a light signal to go from one observer to another and back again is constant in time, or equivalently, that the rod distance between the two observers is constant in time. Any frame of reference can therefore also be called a *rigid* collection of observers. We therefore note that a general frame of reference is *not* the same as a set of coordinates; the latter usually is *not* rigid. In the special case that we have chosen the coordinate system in such a way that all the rigidly connected observers have constant coordinate values, we speak of a *rigid coordinate system*. Obviously, these are the most useful to describe accelerating frames of reference.*

Ref. 273 Note that if two observers both move with a velocity **v**, as measured in some *inertial* frame, they observe that they are at rest with respect to each other *only* if this velocity is

Challenge 631 ny *constant*. Again we find, as above, that two persons tied to each other by a rope, and at a distance such that the rope is under tension, will see the rope break (or hang loose) if they accelerate together to (or decelerate from) relativistic speeds in precisely the same way. Relativistic acceleration requires careful thinking.

An observer who always *feels* the *same* force on his body is called *uniformly* accelerating. More precisely, a uniformly accelerating observer Ω is thus an observer whose acceleration at every moment, measured by the inertial frame with respect to which the observer is at rest *at that moment*, always has the same value **B**. It is important to note that uniform acceleration is *not* uniformly accelerating when always observed from the *same* inertial frame K. This is an important difference with respect to the Galilean case.

For uniformly accelerated motion in the sense just defined, we need

$$\mathbf{B} \cdot \mathbf{B} = -g^2 \tag{181}$$

Ref. 274 where g is a constant independent of t. The simplest case is uniformly accelerating motion that is also *rectilinear*, i.e. for which the acceleration **a** is parallel to **v** at one instant of time

Challenge 632 ny and (therefore) for all other times as well. In this case we can write, using three-vectors,

$$\gamma^3 \mathbf{a} = \mathbf{g} \quad \text{or} \quad \frac{\mathrm{d}\gamma\mathbf{v}}{\mathrm{d}t} = \mathbf{g} \; . \tag{182}$$

Ref. 272 * There are essentially only two other types of rigid coordinate frames, apart from the inertial frames:

- the frame $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 - c^2\mathrm{d}t^2(1 + g_k x_k/c^2)^2$ with arbitrary, but constant acceleration of the origin. The acceleration is $\mathbf{a} = -\mathbf{g}(1 + \mathbf{g}x/c^2)$;
- the uniformly rotating frame $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + 2\omega(-y\,\mathrm{d}x + x\,\mathrm{d}y)\mathrm{d}t - (1 - r^2\omega^2/c^2)\mathrm{d}t$. Here the z-axis is the rotation axis, and $r^2 = x^2 + y^2$.

Taking the direction we are talking about to be the x-coordinate, and solving for $v(t)$, we get

$$v = \frac{gt}{\sqrt{1+\frac{g^2t^2}{c^2}}} , \tag{183}$$

where it was assumed that $v_0 = 0$. We note that for small times we get $v = gt$ and for large times $v = c$, both as expected. The momentum of the Greek observer increases linearly with time, again as expected. Integrating, we find that the accelerated observer Ω moves along the path

Challenge 633 ny

$$x(t) = \frac{c^2}{g}\sqrt{1+\frac{g^2t^2}{c^2}} , \tag{184}$$

where it was assumed that $x_0 = c^2/g$, in order to keep the expression simple. Because of this result, visualized in Figure 165, a rectilinearly and uniformly accelerating observer is said to undergo *hyperbolic* motion. For small times, the world-line reduces to the usual $x = gt^2/2 + x_0$, whereas for large times the result is $x = ct$, as expected. The motion is thus uniformly accelerated only for the moving body itself, *not* for an outside observer.

The proper time τ of the accelerated observer is related to the time t of the inertial frame in the usual way by $\mathrm{d}t = \gamma \mathrm{d}\tau$. Using the expression for the velocity $v(t)$ of equation (183) we get*

Ref. 274, Ref. 275

$$t = \frac{c}{g}\sinh\frac{g\tau}{c} \quad \text{and} \quad x = \frac{c^2}{g}\cosh\frac{g\tau}{c} \tag{185}$$

for the relationship between proper time τ and the time t and the position x measured by the external, inertial Roman observer. We will encounter this relation again during the study of black holes.

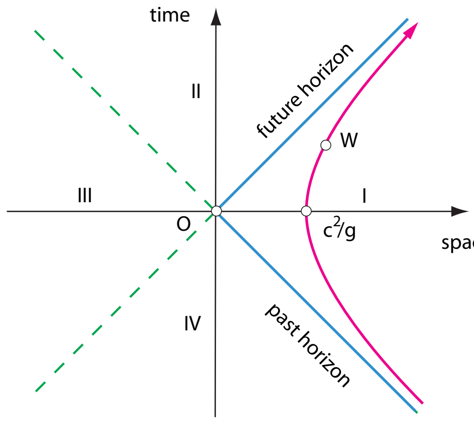


Figure 165 The hyperbolic motion of an rectilinearly, uniformly accelerating observer and its event horizons

Does all this sound boring? Just imagine accelerating on a motor bike at $g = 10\,\mathrm{m/s^2}$ for the proper time τ of 25 years. That would bring you beyond the end of the known universe! Isn't that worth a try? Unfortunately, neither motor bikes nor missiles that accelerate like this exist, as their fuel tank would have to be enormous. Can you confirm this even for the most optimistic case?

Challenge 634 n

Ref. 276

* Use your favourite mathematical formula collection to deduce this. The abbreviations $\sinh y = (e^y - e^{-y})/2$ and $\cosh y = (e^y + e^{-y})/2$ defining the *hyperbolic sine* and the *hyperbolic cosine* imply that $\int \mathrm{d}y/\sqrt{y^2+a^2} = \operatorname{arsinh} y/a = \operatorname{Arsh} y/a = \ln(y+\sqrt{y^2+a^2})$.

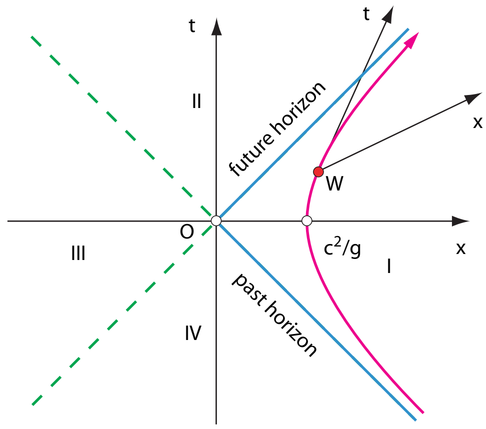


Figure 166 Do accelerated objects depart from inertial ones?

For uniform acceleration, the coordinates transform as

$$
\begin{aligned}
t &= \left(\frac{c}{g}+\frac{\xi}{c}\right)\sinh\frac{g\tau}{c}\\
x &= \left(\frac{c^2}{g}+\xi\right)\cosh\frac{g\tau}{c}\\
y &= \upsilon\\
z &= \zeta\,,
\end{aligned}
\qquad (186)
$$

where τ now is the time coordinate in the Greek frame. We note also that the space-time interval $\mathrm{d}\sigma$ becomes

$$\mathrm{d}\sigma^2 = (1+g\xi/c^2)^2c^2\mathrm{d}\tau^2 - \mathrm{d}\xi^2 - \mathrm{d}\upsilon^2 - \mathrm{d}\zeta^2 = c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \qquad (187)$$

and since for $\mathrm{d}\tau = 0$ distances are given by Pythagoras' theorem, the Greek reference frame is indeed rigid. Ref. 277

After this forest of formulae, let's tackle a simple question. The Roman observer O sees the Greek observer Ω departing with acceleration g, moving further and further away, following equation (184). What does the Greek observer say about his Roman colleague? With all the experience we have now, that is easy. At each point of his trajectory the Greek observer sees that O has the coordinate $\tau = 0$ (can you confirm this?), which means that the distance to the Roman observer, as seen by Greek one, is the same as the space-time interval OΩ. Using expression (184) this turns out to be

Challenge 635 ny

Ref. 278

$$d_{\mathrm{O}\Omega} = \sqrt{\xi^2} = \sqrt{x^2 - c^2t^2} = c^2/g\,, \qquad (188)$$

which, surprisingly enough, is constant in time! In other words, the Greek observer will observe that he stays at a constant distance from the Roman one, in complete contrast to what the Roman observer says. Take your time to check this strange result in some other

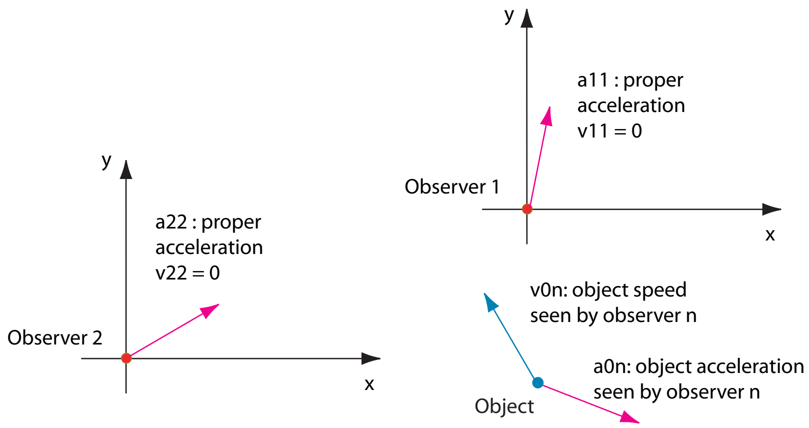


Figure 167 The definitions necessary to deduce the addition theorem for accelerations

way. We will need it again later on, to explain why the Earth does not explode. (Are you able to guess the relationship to this issue?)

Challenge 636 ny

The addition theorem for accelerations is more complex than for velocities. The best explanation was published by Mishra. If we call a_{nm} the acceleration of the system n by observer m, the addition theorem for accelerations asks for the way to express the object acceleration a_{01} as function of the value a_{02} measured by the other observer, the relative acceleration a_{12} and the proper acceleration a_{22} of the other observer. Despite the situation shown in the figure (for clarity reasons) we only study one-dimensional situations, where all observers and the objects move along one direction. (For clarity, we also write $v_{11} = v$ and $v_{02} = u$.) In Galilean physics we have the general connection

Ref. 279
Challenge 637 e

$$a_{01} = a_{02} - a_{12} + a_{22} \tag{189}$$

because accelerations behave simply. In special relativity, one gets

$$a_{01} = a_{02}\frac{(1-v^2/c^2)^{3/2}}{(1-uv/c^2)^3} - a_{12}\frac{(1-u^2/c^2)(1-v^2/c^2)^{-1/2}}{(1-uv/c^2)^2} + a_{22}\frac{(1-u^2/c^2)(1-v^2/c^2)^{3/2}}{(1-uv/c^2)^3} \tag{190}$$

You might enjoy checking the expression.

Are you able to state how the acceleration ratio enters the definition of mass in special relativity?

Challenge 638 ny
Page 284
Challenge 639 ny

Event horizons

The surprises of accelerated motion are not finished yet. Of special interest is the trajectory, in the rigidly accelerated frame coordinates ξ and τ, of an object located at the departure point $x = x_0 = c^2/g$ at all times t. One gets the two relations*

Challenge 640 ny

$$\xi = -\frac{c^2}{g}(1 - \operatorname{sech}\frac{g\tau}{c})$$
$$\mathrm{d}\xi/\mathrm{d}\tau = -c \operatorname{sech}\frac{g\tau}{c} \tanh\frac{g\tau}{c} . \qquad (192)$$

These equations are strange. It is clear that for large times τ the coordinate ξ approaches the limit value $-c^2/g$ and that $\mathrm{d}\xi/\mathrm{d}\tau$ approaches zero. The situation is similar to a car accelerating away from a woman standing on a long road. Seen from the car, the woman moves away; however, after a while, the only thing one notices is that she is slowly approaching the horizon. In Galilean physics, both the car driver and the woman on the road see the other person approaching each other's horizon; in special relativity, only the accelerated observer makes this observation.

Studying a graph of the situation confirms the result. In Figure 166 we can see that light emitted from any event in regions II and III cannot reach the Greek observer. Those events are hidden from him and cannot be observed. Strangely enough, however, light from the Greek observer *can* reach region II. The boundary between the part of space-time that can be observed and that which cannot is called the *event horizon*. In relativity, event horizons act like one-way gates for light and for any other signal. For completeness, the graph also shows the past event horizon. Can you confirm that event horizons are

Challenge 641 ny *black*?

In summary, not all events observed in an inertial frame of reference can be observed in a uniformly accelerating frame of reference. Uniformly accelerating frames of reference produce event horizons at a distance $-c^2/g$. For example, a person who is standing can never see further than this distance below his feet.

By the way, is it true that a light beam *cannot* catch up with an observer in hyperbolic

Challenge 642 n motion, if the observer has a sufficient distance advantage at the start?

Here is a more advanced challenge that prepares for general relativity. What is the

Challenge 643 ny *shape* of the horizon seen by a uniformly accelerated observer?

Acceleration changes colours

We saw above that a moving receiver sees different colours from the sender. This colour shift or Doppler effect was discussed above for inertial motion only. For accelerating frames the situation is even stranger: sender S and receiver R do not agree on colours even

Ref. 274, Ref. 280 if they are at *rest* with respect to each other. Indeed, if light is emitted in the direction of the acceleration, the expression for the space-time interval gives

$$\mathrm{d}\sigma^2 = \left(1 + \frac{g_0 x}{c^2}\right)^2 c^2 \mathrm{d}t^2 \qquad (193)$$

* The functions appearing above, the *hyperbolic secans* and the *hyperbolic tangens*, are defined using the expressions from the footnote on page 304:

$$\operatorname{sech} y = \frac{1}{\cosh y} \quad \text{and} \quad \tanh y = \frac{\sinh y}{\cosh y} . \qquad (191)$$

in which g_0 is the proper acceleration of an observer located at $x = 0$. We can deduce in
Challenge 644 ny a straightforward way that

$$\frac{f_R}{f_S} = 1 - \frac{g_R h}{c^2} = \frac{1}{\left(1 + \frac{g_S h}{c^2}\right)} \tag{194}$$

where h is the rod distance between the source and the receiver, and where $g_S = g_0/(1 + g_0 x_S/c^2)$ and $g_R = g_0/(1 + g_o x_R/c^2)$ are the proper accelerations measured at the x-coordinates of the source and at the detector. In short, the frequency of light decreases when light moves in the direction of acceleration. By the way, does this have an effect on
Challenge 645 n the colour of trees along their vertical extension?

The formula usually given, namely

$$\frac{f_R}{f_S} = 1 - \frac{gh}{c^2} \,, \tag{195}$$

is only correct to first approximation, and not exactly what was just found. In accelerated frames of reference, we have to be careful with the meaning of every quantity used. For everyday accelerations, however, the differences between the two formulae are negligible.
Challenge 646 ny Are you able to confirm this?

Can light move faster than c?

What speed of light is measured by an accelerating observer? Using expression (195) above, an accelerated observer deduces that

$$v_{\text{light}} = c\,(1 + \frac{gh}{c^2}) \tag{196}$$

which is higher than c in the case when light moves in front or 'above' him, and lower than c for light moving behind or 'below' him. This strange result concerning the speed of light follows from a basic property of any accelerating frame of reference. In such a frame, even though all observers are at rest with respect to each other, clocks do *not* remain
Ref. 281 synchronized. The change of the speed of light has also been confirmed by experiment. In other words, the speed of light is only constant when it is defined as $c = \mathrm{d}x/\mathrm{d}t$, *and* if $\mathrm{d}x$ and $\mathrm{d}t$ are measured with a ruler located at a point *inside* the interval $\mathrm{d}x$ and a clock read off *during* an instant inside the interval $\mathrm{d}t$. If the speed of light is defined as $\Delta x/\Delta t$, or if the ruler defining distances or the clock measuring times is located away from the propagating light, the speed of light comes out to be different from c for accelerating observers! This is the same effect you can experience when you turn around your vertcial axis at night: the star velocities you observe are much higher than the speed of light.

Note that this result does not imply that signals or energy can be moved faster than c,
Challenge 647 n as you may want to check for yourself.

In fact, all these difficulties are only noticeable for distances l that do not obey the relation $l \ll c^2/a$. This means that for an acceleration of $9.5\,\mathrm{m/s^2}$, about that of free fall, distances would have to be of the order of one light year, or $9.5 \cdot 10^{12}\,\mathrm{km}$, in order to observe any sizable effects. In short, *c is the speed of light relative to* nearby *matter only.*

Challenge 648 n By the way, everyday gravity is equivalent to a constant acceleration. Why then don't distant objects, such as stars, move faster than light following expression (196)?

What is the speed of light?

We have seen that the speed of light, as usually defined, is given by c only if either the observer is inertial or the observer measures the speed of light passing nearby, instead of light passing at a distance. In short, the speed of light has to be measured locally. But this request does not eliminate all subtleties.

An additional point is often forgotten. Usually, length is measured by the time it takes light to travel. In such a case the speed of light will obviously be constant. However, how does one check the constancy in the present case? One needs to eliminate length measurements. The simplest way to do this is to reflect light from a mirror. The constancy of the speed of light implies that if light goes up and down a short straight line, then the clocks at the two ends measure times given by

$$t_3 - t_1 = 2\,(t_2 - t_1)\ . \tag{197}$$

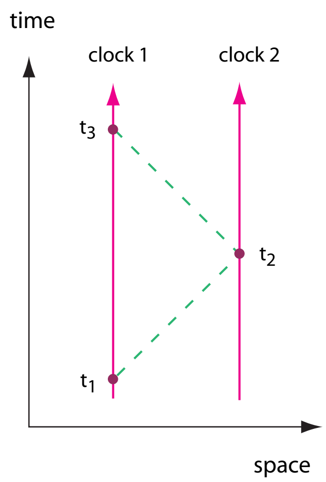


Figure 168 Clocks and the measurement of the speed of light as two-way velocity

Here it was assumed that the clocks were synchronised according to the prescription on page 272. If the factor were not exactly two, the speed of light would not be constant. In fact, all experiments so far have yielded a factor of two within measurement errors. Ref. 282

This result is sometimes expressed by saying that it is impossible to measure the *one-way velocity of light*; only the *two-way* velocity of light is measurable. Do you agree? Challenge 649 n

Limits on the length of solid bodies

An everyday solid object breaks when some part of it moves with more than the speed of sound c of that material with respect to some other part.* For example, when an object hits the floor, its front end is stopped within a distance d; therefore the object breaks at the latest when

$$\frac{v^2}{c^2} \geqslant \frac{2d}{l}\ . \tag{198}$$

We see that we can avoid the breaking of fragile objects by packing them into foam rubber – which increases the stopping distance – of roughly the same thickness as the object's size. This may explain why boxes containing presents are usually so much larger than their contents!

* For glass and metals the (longitudinal) *speed of sound* is about 5.9 km/s for glass, iron or steel, and 4.5 km/s for gold; for lead about 2 km/s. Other sound speeds are given on page 190.

The fracture limit can also be written in a different way. To avoid breaking, the acceleration a of a solid body with length l must follow

$$l a < c^2 , \tag{199}$$

where c is the speed of sound, which is the speed limit for the material parts of solids. Let
Ref. 283 us repeat the argument in relativity, introducing the speed of light instead of that of sound. Imagine accelerating the front of a *solid* body with some *proper* acceleration a. The back end cannot move with an acceleration α equal or larger than infinity, or if one prefers, it
Challenge 650 n cannot move with more than the speed of light. A quick check shows that therefore the length l of a solid body must obey

$$l \alpha < c^2/2 , \tag{200}$$

where c is now the speed of light. The speed of light thus limits the size of solid bodies. For example, for 9.8 $\mathrm{m/s^2}$, the acceleration of a quality motor bike, this expression gives a length limit of 9.2 Pm, about a light year. Not a big restriction; most motor bikes are shorter.

However, there are other, more interesting situations. The highest accelerations achievable today are produced in particle accelerators. Atomic nuclei have a size of a few 1 fm.
Challenge 651 ny Are you able to deduce at which energies they break when smashed together in an accelerator? In fact, inside a nucleus, the nucleons move with accelerations of the order of $v^2/r \approx \hbar^2/m^2 r^3 \approx 10^{31}\ \mathrm{m/s^2}$; this is one of the highest values found in nature.

Note that Galilean physics and relativity produce a similar conclusion: a limiting speed, be it that of sound or that of light, makes it impossible for solid bodies to be *rigid*. When we push one end of a body, the other end always moves a little bit later.

What does this mean for the size of elementary particles? Take two electrons at a distance d, and call their size l. The acceleration due to electrostatic repulsion then leads to
Challenge 652 ny an upper limit for their size given by

$$l < \frac{4\pi\varepsilon_0 c^2 d^2 m}{e^2} . \tag{201}$$

The nearer electrons can get, the smaller they must be. The present experimental limit shows that the size is smaller than 10^{-19} m. Can electrons be exactly point-like? We will come back to this issue during the study of general relativity and quantum theory.

Special relativity in four sentences

This section of our ascent of Motion Mountain is rapidly summarized.

- All (free floating) observers find that there is a unique, perfect velocity in nature, namely a common maximum energy velocity, which is realized by massless radiation such as light or radio signals, but cannot be achieved by material systems.
- Therefore, even though space-time is the same for every observer, times and lengths vary from one observer to another, as described by the Lorentz transformations (112) and (113), and as confirmed by experiment.

▪ Collisions show that a maximum speed implies that mass is concentrated energy, and that the total energy of a body is given by $E = \gamma mc^2$, as again confirmed by experiment.

▪ Applied to accelerated objects, these results lead to numerous counter-intuitive consequences, such as the twin paradox, the appearance of event horizons and the appearance of short-lived tachyons in collisions.

In summary, special relativity shows that motion, though limited in speed, is relative, defined using the propagation of light, conserved, reversible and deterministic.

Could the speed of light vary?

For massless light, the speed of light is the limit speed. Assuming that light is indeed exactly massless, could the speed of light still change from place to place or as time goes by? This tricky question still makes a fool out of many physicists. On first sight, the answer is a loud 'Yes, of course! Just have a look to what happens when the value of c is changed in formulae.' (In fact, there are even attempts to build 'variable speed of light theories'.) However, this often heard statement is wrong.

Since the speed of light enters our definition of time and space, it thus enters, even if we do not notice it, the construction of all rulers, all measurement standards and all measuring set-ups. Therefore there is *no way* to detect whether the value actually varies. No imaginable experiment could detect a variation of the limit speed, as the limit speed is the basis for all measurements. 'That is intellectual cruelty!', you might say. 'All experiments show that the speed of light is invariant; we had to swallow one counter-intuitive result after the other to accept the constancy of the speed of light, and now we are even supposed to admit that there is no other choice?' Yes, we are. That is the irony of progress in physics. The observer invariance of the speed of light is counter-intuitive and astonishing when compared to the lack of observer invariance of everyday, Galilean speeds. But had we taken into account that every speed measurement always is – whether we like it or not – a comparison with the speed of light, we would not have been astonished by the constancy of the speed of light; we would have been astonished by the strange way small speeds behave.

Challenge 653 n

In short, there is no way, in principle, to check the invariance of a standard. To put it in other words, the most counter-intuitive aspect of relativity is not the invariance of c; the most counter-intuitive aspect is the disappearance of c from the formulae of everyday motion.

What happens near the speed of light?

If one approaches the speed of light, the Lorentz transformation expression diverges. A division by zero is impossible; indeed, motion of masses or observers at the speed of light are impossible. However, this is only half the story; there are additional effects.

No observable actually diverges in nature. Approaching the speed of light as much as possible, one always finds that special relativity is not correct any more. At extremely large Lorentz contractions, there is no way to ignore the curvature of space-time; indeed, gravitation has to be taken into account in those cases. Near horizons, there is no way to ignore the fluctuations of speed and position; quantum theory has to be taken into account there. The exploration of these two limitations define the next stages of our ascent of motion mountain.

At the start of our adventure, during our exploration of Galilean physics, once we had defined the basic concepts of velocity, space and time, we turned our attention to gravitation. The invariance of the speed of light has forced us to change these basic concepts; we now return to study gravitation in the light of this invariance.

Bibliography

213 Aristotle, The sense, VI, 446 a 26. Cited in Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 157, 1991. Cited on page 249.

214 The history of the measurement of the velocity of light can be found in chapter 19 of the text by Francis A. Jenkins & Harvey E. White, *Fundamentals of Optics*, McGraw-Hill, New York, 1957. Cited on page 250.

215 On the way to perform such measurements, see Sydney G. Brewer, *Do-it-yourself Astronomy*, Edinborough University Press, 1988. Kepler himself never measured the distances of planets to the Sun, but only *ratios* of planetary distances. The parallax of the Sun from two points of the Earth is at most 8.79 ″; it has been measured already in the eighteenth century.

The measurement of Moon distance is now performed with laser beams fired onto reflectors deposited there during the Apollo and Lunakhod missions. Distance measurements are done by time of flight measurements of 35 ps laser pulses. See the http://www.csr.utexas.edu/mlrs and http://ilrs.gsfc.nasa.gov websites. Cited on page 251.

216 The art to make pictures of light pulses with a usual photographic camera, without any electronics, is told by M.A. Duguay & A.T. Mattick, Ultrahigh speed photography of picosecond light pulses and echoes, *Applied Optics* **10**, 1971, pp. 2162–2170. The picture on page 252 is also taken from there. Cited on page 252.

217 You can learn the basics of special relativity with the help of the web, without referring to any book, by using the http://physics.syr.edu/research/relativity/RELATIVITY.html web page as a starting point. This page mentions most English language relativity pages available on the web. Links in other languages can be found with search engines. Cited on page 254.

218 The constancy of light speed deduced from gamma ray bursts shows that the speed of light does not depend on the lamp speed to within one part in 10^{20}, as shown by K. Brecher, *Bulletin of the American Physical Society* **45**, 2000. He assumed that both sides of the burster emit light. The large speed difference and the pulse sharpness then yield the limit. See also his older paper K. Brecher, Is the speed of light independent of the source?, *Physics Letters* **39**, pp. 1051–1054, Errata 1236, 1977. Measuring the light speed from rapidly moving stars is another way. Some of these experiments are not completely watertight however; there is a competing theory of electrodynamics, by Ritz, which maintains that the speed of light is c only when measured with respect to the source. The light from stars however passes through the atmosphere, and its speed thus might be reduced to c.

The famous experiment with light emitted from rapid pions at CERN does not suffer from this criticism. It is described in T. Alväger, J.M. Bailey, F.J.M. Farley, J. Kjellman & I. Wallin, Test of the second postulate of relativity in the GeV region, *Physics Letters* **12**, pp. 260–262, 1964. See also T. Alväger & al., Velocity of high-energy gamma rays, *Arkiv* *för Fysik* **31**, pp. 145–157, 1965.

Another precise experiment at extreme speeds is described by G.R. Kalbfleisch, N. Baggett, E.C. Fowler & J. Alspector, Experimental comparison of neutrino, antineutrino, and muon velocities, *Physical Review Letters* **43**, pp. 1361–1364, 1979. Cited on page 254.

219 See e.g. C. Will, *Theory and Experiment in Gravitational Physics*, Revised edition, Cambridge University Press, 1993. Cited on pages 254 and 255.

220 B.E. Schaefer, Severe limits on variations of the speed of light with frequency, *Physical Review Letters* **82**, pp. 4964–4966, 21 June 1999. Cited on page 254.

221 The start of the modern theory of relativity is the famous paper by Albert Einstein, Zur Elektrodynamik bewegter Körper, *Annalen der Physik* **17**, pp. 891–921, 1905. It still well worth reading, and every physicist should have done so. The same ican be said of the famous paper, written probably after he heard of Olinto De Pretto's idea, found in Albert Einstein, Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?, *Annalen der Physik* **18**, pp. 639–641, 1905.See also the review Albert Einstein, Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen, *Jahrbuch der Radioaktivität und Elektronik* **4**, pp. 411–462, 1907. They are now available in many languages. A later, unpublished review is available in facsimile and with an English translation as Albert Einstein, Hanoch Gutfreund, ed., Einstein's 1912 Manuscript on the Theory of Relativity, George Braziller, 2004. Cited on pages 254, 256, and 290.

222 Albert Einstein, *Mein Weltbild*, edited by Carl Selig, Ullstein Verlag, 1998. Cited on page 255.

223 Jean van Bladel, *Relativity and Engineering*, Springer 1984. Cited on page 255.

224 Albrecht Fölsing, *Albert Einstein - eine Biographie*, 1993, p. 237. Cited on page 255.

225 R.J. Kennedy & E.M. Thorndike, Experimental establishment of the relativity of time, *Physical Review* **42**, pp. 400–418, 1932. See also H.E. Ives & G.R. Stilwell, An experimental study of the rate of a moving atomic clock, *Journal of the Optical Society of America* **28**, pp. 215–226, 1938, and **31**, pp. 369–374, 1941, For a modern, high precision version, see C. Braxmeier, H. Müller, O. Pradl, J. Mlynek, A. Peters & S. Schiller, New tests of relativity using a cryogenic optical resonator, *Physical Review Letters* **88**, p. 010401, 2002. Cited on page 256.

226 Edwin F. Taylor & John A. Wheeler, *Spacetime Physics – Introduction to Special Relativity*, second edition, Freeman, 1992. Another excellent book is Wolfgang Rindler, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001. See also Nick M.J. Woodhouse, *Special Relativity*, Springer, 2003. Cited on pages 256 and 297.

227 The art of explaining special relativity by drawing a few lines on paper is due to Hermann Bondi, *Relativity and Common Sense: A New Approach to Einstein*, Dover , New York, 1980. See also Dierck-Ekkehard Liebscher, *Relativitätstheorie mit Zirkel und Lineal*, Akademie-Verlag Berlin, 1991. Cited on page 256.

228 Rod S. Lakes, Experimental limits on the photon mass and cosmic vector potential, *Physical Review Letters* **80**, pp. 1826–1829, 1998. The speed of light is independent of frequency within a factor of $6 \cdot 10^{-21}$, as was shown from gamma ray studies by B.E. Schaefer, Severe limits on variations of the speed of light with frequency, *Physical Review Letters* **82**, pp. 4964–4966, 1999. Cited on page 258.

229 An overview of previous results is given in Yuan Zhong Zhang, *Special Relativity and its Experimental Foundations*, World Scientific, 1998. Cited on pages 258, 272, 315, and 317.

230 R.W. McGowan & D.M. Giltner, New measurement of the relativistic Doppler shift in neon, *Physical Review Letters* **70**, pp. 251–254, 1993. Cited on page 259.

231 The present record for clock synchronization seems to be 1 ps for two clocks distant 3 km from each other. See A. Valencia, G. Scarcelli & Y. Shih, Distant clock synchronization using entangled photon pairs, *Applied Physics Letters* **85**, pp. 2655–2657, 2004, or http://www.arxiv.org/absquant-ph/0407204. Cited on page 260.

232 J. Frenkel & T. Kontorowa, Über die Theorie der plastischen Verformung, *Physikalische Zeitschrift der Sowietunion* **13**, p. 1, 1938. F.C. Frank, On the equations of motion of crystal dislocations, *Proceedings of the Physical Society A* **62**, pp. 131–134, 1949, J. Eshelby, Uniformly moving dislocations, *Proceedings of the Physical Society A* **62**, pp. 307–314, 1949. See also G. Leibfried & H. Dietze, *Zeitschrift für Physik* **126**, p. 790, 1949. A general introduction can be found in A. Seeger & P. Schiller, Kinks in dislocations lines and their effects in internal friction in crystals, *Physical Acoustics* **3A**, W.P. Mason, ed., Academic Press, 1966. See also the textbooks by Frank R.N. Nabarro, *Theory of Crystal Dislocations*, Oxford University Press, 1967, or J.P. Hirth & J. Lothe, *Theory of Dislocations*, McGraw Hills Book Company, 1968. Cited on page 260.

233 The beautiful graph is taken from Z.G.T. Guiragossian, G.B. Rothbart, M.R. Yearian, R. Gearhart & J.J. Murray, Relative velocity measurements of electrons and gamma rays at 15 GeV, *Physical Review Letters* **34**, pp. 335–338, 1975. Cited on page 261.

234 To know more about the best known crackpots, their ideas, e-mail to majordomo@zikzak.net with the one line body 'subscribe psychoceramics'. Cited on page 261.

235 The speed of neutrinos is the same as that of light within 9 decimal digits, since neutrinos and light arrived together, within 12 seconds, after a trip of 170 000 light years from a supernova explosion. This is explained by Leo Stodolsky, The speed of light and the speed of neutrinos, *Physics Letters B* **201**, p. 353, 1983. An observation of a small mass for the neutrino has been published by the Japanese Super-Kamiokande collaboration, in Y. Fukuda & al., Evidence for oscillation of atmospheric neutrinos, *Physical Review Letters* **81**, pp. 1562 –1567, 1998. The newer results published by the Canadian Sudbury Neutrino Observatory, as Q.R. Ahmad & al., Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory, *Physical Review Letters* **89**, p. 011301, 2002, also confirm that neutrinos have a mass in the 1 eV region. Cited on page 262.

236 B. Rothenstein & G. Eckstein, Lorentz transformations directly from the speed of light, *American Journal of Physics* **63**, p. 1150, 1995. See also the comment by E. Kapuścik, Comment on "Lorentz transformations directly from the speed of light," by B. Rothenstein and G. Eckstein, *American Journal of Physics* **65**, p. 1210, 1997. Cited on page 263.

237 See e.g. the 1922 lectures by Lorentz at Caltech, published as H.A. Lorentz, *Problems of Modern Physics*, edited by H. Bateman, Ginn and Company, 1927, page 99. Cited on page 263.

238 A.A. Michelson & E.W. Morley, On the relative motion of the Earth and the luminiferous ether, *American Journal of Science (3rd series)* **34**, pp. 333–345, 1887. Michelson published many other papers on the topic after this one. Cited on page 263.

239 H.A. Lorentz, De relative beweging van de aarde en dem aether, *Amst. Versl.* **1**, p. 74, 1892, and also H.A. Lorentz, Electromagnetic phenomena in a system moving with any velocity smaller than that of light, *Amst. Proc.* **6**, p. 809, 1904, or *Amst. Versl.* **12**, p. 986, 1904, Cited on page 266.

240 A general refutation of such proposals is discussed by S.R. Mainwaring & G.E. Stedman, Accelerated clock principles, *Physical Review A* **47**, pp. 3611–3619, 1993. Already exper-

iments on *muons* at CERN in 1968 showed that *accelerations* of up to 10^{20} m/s^2 have no effect, as explained by D.H. PERKINS, *Introduction to High Energy Physics*, Addison-Wesley, 1972, or by J. BAILEY & al., *Il Nuovo Cimento* **9A**, p. 369, 1972. Cited on page 266.

241 W. RINDLER, General relativity before special relativity: an unconventional overview of relativity theory, *American Journal of Physics* **62**, pp. 887–893, 1994. Cited on page 267.

242 STEVEN K. BLAU, Would a topology change allow Ms. Bright to travel backward in time?, *American Journal of Physics* **66**, pp. 179–185, 1998. Cited on page 269.

243 On the 'proper' formulation of relativity, see for example ... Cited on page 270.

244 The simple experiment to take a precise clock on a plane, fly it around the world and then compare it with an identical one left in place was first performed by J.C. HAFELE & R.E. KEATING, Around-the-world atomic clocks: predicted relativistic time gains, *Science* **177**, pp. 166–167, and Around-the-world atomic clocks: observed relativistic time gains, pp. 168–170, 14 July 1972. See also reference Ref. 229. Cited on page 270.

245 A readable introduction into the change of time with observers and relativity in general is ROMAN U. SEXL & HERBERT KURT SCHMIDT, *Raum-Zeit-Relativität*, 2. Auflage, Vieweg & Sohn, Braunschweig, 1991. Cited on page 270.

246 PAUL J. NAHIN, *Time Machines – Time Travel in Physics, Metaphysics and Science Fiction*, Springer Verlag and AIP Press, second edition, 1999. Cited on page 270.

247 The first muon experiment was B. ROSSI & D.B. HALL, Variation of the rate of decay of mesotrons with momentum, *Physical Review* **59**, pp. 223–228, 1941. Cited on page 271.

248 A. HARVEY & E. SCHUCKING, A small puzzle from 1905, *Physics Today* pp. 34–36, March 2005. Cited on page 272.

249 Most famous is the result that moving muons stay younger, as shown for example by D.H. FRISCH & J.B. SMITH, Measurement of the relativistic time dilation using μ-mesons, *American Journal of Physics* **31**, pp. 342–335, 1963. Cited on page 270.

250 W. RINDLER, Length contraction paradox, *American Journal of Physics* **29**, 1961, pp. 365–366, For a variation without gravity, see R. SHAW, Length contraction paradox, *American Journal of Physics* **30**, p. 72, 1962. Cited on page 273.

251 H. VAN LINTEL & C. GRUBER, The rod and hole paradox re-examined, *European Journal of Physics* **26**, pp. 19–23, 2005. Cited on page 273.

252 This situation is discussed by G.P. SASTRY, Is length contraction paradoxical?, *American Journal of Physics* **55**, 1987, pp. 943–946, It also contains an extensive literature list of all variants of length contraction paradoxes. Cited on page 274.

253 S.P. BOUGHN, The case of the identically accelerated twins, *American Journal of Physics* **57**, pp. 791–793, 1989. Cited on pages 274 and 277.

254 J.M. SUPPLEE, Relativistic buoyancy, *American Journal of Physics 57* **1**, pp. 75–77, January 1989. See also G.E.A. MATSAS, Relativistic Arquimedes law for fast moving bodies and the general-relativistic resolution of the 'submarine paradox', *Physical Review D* **68**, p. 027701, 2003, or http://www.arxiv.org/abs/gr-qc/0305106. Cited on page 274.

255 G.R. RYBICKI, Speed limit on walking, *American Journal of Physics* **59**, pp. 368–369, 1991. Cited on page 278.

256 The slowness of the speed of light inside stars is explained e.g. by ... Cited on page 279.

257 The first examples of such astronomical observations were provided by A.R. WHITNEY & al., Quasars revisited: rapid time variations observed via very-long-baseline interferometry, *Science* **173**, pp. 225–230, 1971, and by M.H. COHEN & al., The small-scale structure of radio galaxies and quasi-stellar sources at 3.8 centimetres, *Astrophysical Journal* **170**,

pp. 207–217, 1971. See also T.J. Pearson, S.C. Unwin, M.H. Cohen, R.P. Linfield, A.C.S. Readhead, G.A. Seielstad, R.S. Simon & R.C. Walker, Superluminal expansion of quasar 3C 273, *Nature* **290**, pp. 365–368, 1981. An overview is given in J.A. Zensus & T.J. Pearson, editors, *Superluminal radio sources*, Cambridge University Press, 1987. Another measurement, using very long baseline interferometry with radiowaves, was shown on the cover of nature: I.F. Mirabel & L.F. Rodriguez, A superluminal source in the galaxy, *Nature* **371**, pp. 46–48, 1994. Cited on page 280.

A more recent example was reported in *Science News* **152**, p. 357, 6 December 1997.

A pedagogical explanation is given by D.C. Gabuzda, The use of quasars in teaching introductory special relativity, *American Journal of Physics* **55**, pp. 214–215, 1987, and by Edwin F. Taylor & John A. Wheeler, *Spacetime Physics – Introduction to Special Relativity*, second edition, Freeman, 1992, pages 89-92. This excellent book was mentioned already in the text.

258 O.M. Bilaniuk & E.C. Sudarshan, Particles beyond the light barrier, *Physics Today* **22**, pp. 43–51, 1969, O.M.P. Bilanuk, V.K. Deshpande & E.C.G. Sudarshan, 'Meta' relativity, *American Journal of Physics* **30**, pp. 718–723, 1962. See also E. Recami, editor, *Tachyons, Monopoles and Related Topics*, North-Holland, Amsterdam, 1978. Cited on page 281.

259 J.P. Costella, B.H.J. McKellar, A.A. Rawlinson & G.J. Stephenson, The Thomas rotation, *American Journal of Physics* **69**, pp. 837–847, 2001. Cited on page 282.

260 See for example S.S. Costa & G.E.A. Matsas, Temperature and relativity, preprint available at http://www.arxiv.org/abs/gr-qc/9505045. Cited on page 282.

261 R.C. Tolman & G.N. Lewis, The principle of relativity and non-Newtonian mechanics, *Philosophical Magazine* **18**, pp. 510–523, 1909, and R.C. Tolman, Non-Newtonian mechanics: the mass of a moving body, *Philosophical Magazine* **23**, pp. 375–380, 1912. Cited on page 283.

262 This information is due to a private communication by Frank DiFilippo; part of the story is given in F. DiFilippo, V. Natarajan, K.R. Boyce & D.E. Pritchard, Accurate atomic masses for fundamental metrology, *Physical Review Letters* **73**, pp. 1481–1484, 1994. These measurements were performed with Penning traps; a review of the possibilities they offer is given by R.C. Thompson, Precision measurement aspects of ion traps, *Measurement Science and Technology* **1**, pp. 93–105, 1990. The most important experimenters in the field of single particle levitation were awarded the Nobel prize in 1989. One of the Nobel prize lectures can be found in W. Paul, Electromagnetic traps for neutral and charged particles, *Reviews of Modern Physics* **62**, pp. 531–540, 1990. Cited on page 287.

263 J.L. Synge, *Relativity: The Special Theory*, North-Holland, 1956, pp. 208–213. More about antiparticles in special relativity can be found in J.P. Costella, B.H.J. McKellar & A.A. Rawlinson, Classical antiparticles, *American Journal of Physics* **65**, pp. 835–841, 1997. See also reference Ref. 278. Cited on page 287.

264 A. Papapetrou, Drehimpuls- und Schwerpunktsatz in der relativistischen Mechanik, *Praktika Acad. Athenes* **14**, p. 540, 1939, A. Papapetrou, Drehimpuls and Schwerpunktsatz in der Diracschen Theorie, *Praktika Acad. Athenes* **15**, p. 404, 1940, M.H.L. Pryce, The mass-centre in the restricted theory of relativity and its connexion with the quantum theory of elementary particle, *Proceedings of the Royal Society in London, A* **195**, pp. 62–81, 1948. Cited on page 290.

265 A jewel among the textbooks on special relativity is the booklet by Ulrich E. Schröder, *Spezielle Relativitätstheorie*, Verlag Harri Deutsch, Thun, 1981. Cited on pages 293 and 295.

266 A readable article showing a photocopy of a letter by Einstein making this point is Lev B. Okun, The concept of mass, *Physics Today* pp. 31–36, June 1989. The topic is not without controversy, as the letters following that article show; they are found in Putting to rest mass misconceptions, *Physics Today* pp. 13–14, as well as pp. 115–117, May 1990. The topic is still source of debates. Cited on page 295.

267 Christian Møller, *The Theory of Relativity*, Clarendon Press, 1952, 1972. This standard text has also been translated in several other languages. Cited on page 295.

268 P. Ehrenfest, Gleichförmige Rotation starrer Körper und Relativitätstheorie, *Physikalische Zeitschrift* **10**, pp. 918–928, 1909. Ehrenfest (incorrectly) suggested that this meant that relativity cannot be correct. A modern summary of the issue can be found in M.L. Ruggiero, The relative space: space measurements on a rotating platform, http://www.arxiv.org/abs/gr-qc/0309020. Cited on page 296.

269 R.J. Low, When moving clocks run fast, *European Journal of Physics* **16**, pp. 228–229, 1995. Cited on pages 300 and 301.

270 Wolfgang Rindler, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001. A beautiful book by one of the masters of the field. Cited on page 301.

271 G. Stephenson & C.W. Kilmister, *Special Relativity for Physicists*, Longmans, London, 1965. See also W.N. Matthews, Relativistic velocity and acceleration transformations from thought experiments, *American Journal of Physics* **73**, pp. 45–51, 2005. Cited on page 302.

272 The impossibility to define rigid coordinate frames for non-uniformly accelerating observers is discussed by Charles Misner, Kip Thorne & John A. Wheeler, *Gravitation*, Freeman, 1973, page 168. Cited on page 303.

273 E.A. Desloge & R.J. Philpott, Uniformly accelerated reference frames in special relativity, *American Journal of Physics* **55**, pp. 252–261, 1987. Cited on page 303.

274 R.H. Good, Uniformly accelerated reference frame and twin paradox, *American Journal of Physics* **50**, pp. 232–238, 1982. Cited on pages 303, 304, and 307.

275 J. Dwayne Hamilton, The uniformly accelerated reference frame, *American Journal of Physics* **46**, pp. 83–89, 1978. Cited on page 304.

276 The best and cheapest mathematical formula collection remains the one by K. Rottmann, *Mathematische Formelsammlung*, BI Hochschultaschenbücher, 1960. Cited on page 304.

277 C.G. Adler & R.W. Brehme, Relativistic solutions to a falling body in a uniform gravitation field, *American Journal of Physics* **59**, pp. 209–213, 1991. Cited on page 305.

278 See for example the excellent lecture notes by D.J. Raymond, A radically modern approach to freshman physics, on the http://www.physics.nmt.edu/~raymond/teaching.html website. Cited on pages 305 and 316.

279 L. Mishra, The relativistic acceleration addition theorem, *Classical and Quantum Gravity* **11**, pp. L97–L102, 1994. Cited on page 306.

280 Edward A. Desloge, The gravitational red-shift in a uniform field, *American Journal of Physics* **58**, pp. 856–858, 1990. Cited on page 307.

281 The propagation delays to be discussed in the chapter on general relativity can be seen as confirmations of this effect. Cited on page 308.

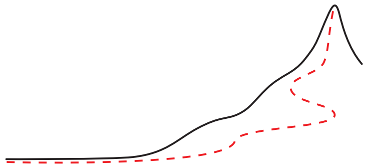
282 The subtleties of the one-way and two-way speed of light will remain a point of discussion for a long time. Many experiments are explained and discussed in reference Ref. 229. Zhang says in his summary on page 171, that the one-way velocity of light is indeed independent of the light source; however, no experiment really shows that it is equal to the two way velocity.

Moreover, most so called 'one-way' experiments are in fact still 'two-way' experiments (see page 150); one can only say that the two-way velocity is anisotropic. Cited on page 309.

283 Edwin F. Taylor & A.P. French, Limitation on proper length in special relativity, *American Journal of Physics* **51**, pp. 889–893, 1983. Cited on page 310.

Chapter III

Gravitation and Relativity



General relativity is easy. Nowadays, it can be made as intuitive as universal gravity and its inverse square law – by using the right approach. The main ideas of general relativity, like those of special relativity, are accessible to secondary-school students. Black holes, gravitational waves, space-time curvature and the limits of the universe can then be understood with as easily as the Doppler effect or the twins paradox.

We will discover that, just as special relativity is based on a maximum speed c, general relativity is based on a maximum force $c^4/4G$ or on a maximum power $c^5/4G$. We first show that all known experimental data are consistent with these limits. In fact, we find that the maximum force and the maximum power are achieved only on insurmountable limit surfaces; these limit surfaces are called *horizons*. We will then be able to deduce the field equations of general relativity. In particular, the existence of a maximum for force or power implies that space-time is curved. It explains why the sky is dark at night, and it shows that the universe is of finite size.

We also discuss the main counter-arguments and paradoxes arising from the limits. The resolutions of the paradoxes clarify why the limits have remained dormant for so long, both in experiments and in teaching.

After this introduction, we will study the effects of relativistic gravity in more detail. In particular, we will study the consequences of space-time curvature for the motions of bodies and of light in our everyday environment. For example, the inverse square law will be modified. (Can you explain why this is necessary in view of what we have learned so far?) Most fascinating of all, we will discover how to move and bend the vacuum. Then we will study the universe at large; finally, we will explore the most extreme form of gravity: black holes.

Challenge 654 n

7. Maximum force: general relativity in one statement

We just saw that the theory of *special* relativity appears when we recognize the speed limit c in nature and take this limit as a basic principle. At the end of the twentieth century it was shown that general relativity can be approached by using a similar basic principle:*

▷ *There is in nature a maximum force:*

$$F \leqslant \frac{c^4}{4G} = 3.0 \cdot 10^{43}\,\mathrm{N}\,. \qquad (202)$$

Ref. 284, Ref. 285, Ref. 286 Ref. 287

* This principle was published in the year 2000 in this text, and independently in a conference proceedings in 2002 by Gary Gibbons. The present author discovered the maximum force in 1998 when searching for a way to derive the results of chapter XI that would be so simple that it would convince even a secondary-school student.

In nature, no force in any muscle, machine or system can exceed this value.

For the curious, the value of the force limit is the energy of a (Schwarzschild) black hole divided by twice its radius. The force limit can be understood intuitively by noting that (Schwarzschild) black holes are the densest bodies possible for a given mass. Since there is a limit to how much a body can be compressed, forces – whether gravitational, electric, centripetal or of any other type – cannot be arbitrary large.

Alternatively, it is possible to use another, equivalent statement as a basic principle:

▷ *There is a maximum power in nature:*

$$P \leqslant \frac{c^5}{4G} = 9.1 \cdot 10^{51}\,\mathrm{W}\ . \qquad (203)$$

No power of any lamp, engine or explosion can exceed this value. The maximum power is realized when a (Schwarzschild) black hole is radiated away in the time that light takes to travel along a length corresponding to its diameter. We will see below precisely what black holes are and why they are connected to these limits.

The existence of a maximum force or power implies the full theory of general relativity. In order to prove the correctness and usefulness of this approach, a sequence of arguments is required. The sequence is the same as for the establishment of the limit speed in special relativity. First of all, we have to gather all observational evidence for the claimed limit. Secondly, in order to establish the limit as a principle of nature, we have to show that general relativity follows from it. Finally, we have to show that the limit applies in all possible and imaginable situations. Any apparent paradoxes will need to be resolved.

These three steps structure this introduction to general relativity. We start the story by explaining the origin of the idea of a limiting value.

The maximum force and power limits

In the nineteenth and twentieth centuries many physicists took pains to avoid the concept of force. Heinrich Hertz made this a guiding principle of his work, and wrote an influential textbook on classical mechanics without ever using the concept. The fathers of quantum theory, who all knew this text, then dropped the term 'force' completely from the vocabulary of microscopic physics. Meanwhile, the concept of 'gravitational force' was eliminated from general relativity by reducing it to a 'pseudo-force'. Force fell out of fashion.

Nevertheless, the maximum force principle does make sense, provided that we visualize it by means of the useful definition: *force is the flow of momentum per unit time*. Momentum cannot be created or destroyed. We use the term 'flow' to remind us that momentum, being a conserved quantity, can only change by inflow or outflow. In other words, change of momentum always takes place through some boundary surface. This fact is of central importance. Whenever we think about force at a point, we mean the momentum 'flowing' through a surface at that point.

The maximum force principle thus boils down to the following: if we imagine any physical surface (and cover it with observers), the integral of momentum flow through the surface (measured by all those observers) never exceeds a certain value. It does not matter how the surface is chosen, as long as it is physical, i.e., as long as we can fix observers* onto it.

This principle imposes a limit on muscles, the effect of hammers, the flow of material, the acceleration of massive bodies, and much more. No system can create, measure or experience a force above the limit. No particle, no galaxy and no bulldozer can exceed it.

The existence of a force limit has an appealing consequence. In nature, forces can be measured. Every measurement is a comparison with a standard. The force limit provides a *natural* unit of force which fits into the system of natural units* that Max Planck derived from c, G and h (or $\hbar$). The maximum force thus provides a standard of force valid in every place and at every instant of time.

The limit value of $c^4/4G$ differs from Planck's proposed unit in two ways. First, the numerical factor is different (Planck had in mind the value c^4/G). Secondly, the force unit is a *limiting* value. In the this respect, the maximum force plays the same role as the Ref. 286 maximum speed. As we will see later on, this limit property is valid for all other Planck Page 985 units as well, once the numerical factors have been properly corrected. The factor 1/4 has no deeper meaning: it is just the value that leads to the correct form of the field equations of general relativity. The factor 1/4 in the limit is also required to recover, in everyday Page 337 situations, the inverse square law of universal gravitation. When the factor is properly taken into account, the maximum force (or power) is simply given by the (corrected) Planck energy divided by the (corrected) Planck length or Planck time.

The expression for the maximum force involves the speed of light c and the gravitational constant G; it thus qualifies as a statement on relativistic gravitation. The fundamental principle of special relativity states that speed v obeys $v \leqslant c$ for all observers. Analogously, the basic principle of general relativity states that in all cases force F and power P obey $F \leqslant c^4/4G$ and $P \leqslant c^5/4G$. It does not matter whether the observer measures the force or power while moving with high velocity relative to the system under observation, during free fall, or while being strongly accelerated. However, we will see that it is essential that the observer records values measured *at his own location* and that the observer is *realistic*, i.e., made of matter and not separated from the system by a horizon. These conditions are the same that must be obeyed by observers measuring velocity in special relativity.

Since physical power is force times speed, and since nature provides a speed limit, the force bound and the power bound are equivalent. We have already seen that force Page 295 and power appear together in the definition of 4-force; we can thus say that the upper bound is valid for every component of a force, as well as for its magnitude. The power bound limits the output of car and motorcycle engines, lamps, lasers, stars, gravitational radiation sources and galaxies. It is equivalent to $1.2 \cdot 10^{49}$ horsepowers. The maximum power principle states that there is no way to move or get rid of energy more quickly than that.

* Observers in general relativity, like in special relativity, are massive physical systems that are small enough so that their influence on the system under observation is negligible.

Page 658 * When Planck discovered the quantum of action, he had also noticed the possibility to define natural units. On a walk with his seven-year-old son in the forest around Berlin, he told him that he had made a discovery as important as the discovery of universal gravity.

The power limit can be understood intuitively by noting that every engine produces *exhausts*, i.e. some matter or energy that is left behind. For a lamp, a star or an evaporating black hole, the exhausts are the emitted radiation; for a car or jet engine they are hot gases; for a water turbine the exhaust is the slowly moving water leaving the turbine; for a rocket it is the matter ejected at its back end; for a photon rocket or an electric motor it is electromagnetic energy. Whenever the power of an engine gets close to the limit value, the exhausts increase dramatically in mass–energy. For extremely high exhaust masses, the gravitational attraction from these exhausts – even if they are only radiation – prevents further acceleration of the engine with respect to them. The maximum power principle thus expresses that there is a built-in braking mechanism in nature; this braking mechanism is gravity.

Yet another, equivalent limit appears when the maximum power is divided by c^2.

▷ *There is a maximum rate of mass change in nature:*

$$\frac{\mathrm{d}m}{\mathrm{d}t} \leqslant \frac{c^3}{4G} = 1.0 \cdot 10^{35}\,\mathrm{kg/s}\;. \tag{204}$$

This bound imposes a limit on pumps, jet engines and fast eaters. Indeed, the rate of flow of water or any other material through tubes is limited. The mass flow limit is obviously equivalent to either the force or the power limit.

The claim of a maximum force, power or mass change in nature seems almost too fantastic to be true. Our first task is therefore to check it empirically as thoroughly as we can.

The experimental evidence

Like the maximum speed principle, the maximum force principle must first of all be checked experimentally. Michelson spent a large part of his research life looking for possible changes in the value of the speed of light. No one has yet dedicated so much effort to testing the maximum force or power. However, it is straightforward to confirm that no experiment, whether microscopic, macroscopic or astronomical, has ever measured force values larger than the stated limit. Many people have claimed to have produced speeds larger than that of light. So far, nobody has ever claimed to have produced a force larger than the limit value.

The large accelerations that particles undergo in collisions inside the Sun, in the most powerful accelerators or in reactions due to cosmic rays correspond to force values much smaller than the force limit. The same is true for neutrons in neutron stars, for quarks inside protons, and for all matter that has been observed to fall towards black holes. Furthermore, the search for space-time singularities, which would allow forces to achieve or exceed the force limit, has been fruitless.

In the astronomical domain, all forces between stars or galaxies are below the limit value, as are the forces in their interior. Not even the interactions between any two halves of the universe exceed the limit, whatever physically sensible division between the two halves is taken. (The meaning of 'physically sensible division' will be defined below; for divisions that are *not* sensible, exceptions to the maximum force claim *can* be constructed.

Page 335

Challenge 655 n You might enjoy searching for such an exception.)

Astronomers have also failed to find any region of space-time whose curvature (a concept to be introduced below) is large enough to allow forces to exceed the force limit. Indeed, none of the numerous recent observations of black holes has brought to light forces larger than the limit value or objects smaller than the corresponding black hole radii. Observations have also failed to find a situation that would allow a rapid observer to observe a force value that exceeds the limit due to the relativistic boost factor.

The power limit can also be checked experimentally. It turns out that the power – or luminosity – of stars, quasars, binary pulsars, gamma ray bursters, galaxies or galaxy clusters can indeed be close to the power limit. However, no violation of the limit has Ref. 288 ever been observed. Even the sum of all light output from all stars in the universe does not exceed the limit. Similarly, even the brightest sources of gravitational waves, merging black holes, do not exceed the power limit. Only the brightness of evaporating black holes in their final phase could equal the limit. But so far, none has ever been observed.

Similarly, all observed mass flow rates are orders of magnitude below the corresponding limit. Even physical systems that are mathematical analogues of black holes – for example, silent acoustical black holes or optical black holes – do not invalidate the force and power limits that hold in the corresponding systems.

The experimental situation is somewhat disappointing. Experiments do not contradict the limit values. But neither do the data do much to confirm them. The reason is the lack of horizons in everyday life and in experimentally accessible systems. The maximum speed at the basis of special relativity is found almost everywhere; maximum force and maximum power are found almost nowhere. Below we will propose some dedicated tests Page 339 of the limits that could be performed in the future.

Deducing general relativity*

In order to establish the maximum force and power limits as fundamental physical principles, it is not sufficient to show that they are consistent with what we observe in nature. It is necessary to show that they imply the complete theory of general relativity. (This section is only for readers who already know the field equations of general relativity. Other Page 327 readers may skip to the next section.)

In order to derive the theory of relativity we need to study those systems that *realize* the limit under scrutiny. In the case of the special theory of relativity, the main system that realizes the limit speed is light. For this reason, light is central to the exploration of special relativity. In the case of general relativity, the systems that realize the limit are less obvious. We note first that a maximum force (or power) cannot be realized throughout a *volume* of space. If this were possible, a simple boost** could transform the force (or power) to a higher value. Therefore, nature can realize maximum force and power only on surfaces, not volumes. In addition, these surfaces must be unattainable. These unattainable surfaces Ref. 286 are basic to general relativity; they are called *horizons*. Maximum force and power only Page 307 appear on horizons. We have encountered horizons in special relativity, where they were

* This section can be skipped at first reading.

** A *boost* was defined in special relativity as a change of viewpoint to a second observer *moving* in relation to the first.

defined as surfaces that impose limits to observation. (Note the contrast with everyday life, where a horizon is only a line, not a surface.) The present definition of a horizon as a surface of maximum force (or power) is equivalent to the definition as a surface beyond which no signal may be received. In both cases, a horizon is a surface beyond which interaction is impossible.

The connection between horizons and the maximum force is a central point of relativistic gravity. It is as important as the connection between light and the maximum speed in special relativity. In special relativity, we showed that the fact that light speed is the maximum speed in nature implies the Lorentz transformations. In general relativity, we will now prove that the maximum force in nature, which we can call the *horizon force*, implies the field equations of general relativity. To achieve this aim, we start with the realization that all horizons have an energy flow across them. The flow depends on the horizon curvature, as we will see. This connection implies that horizons cannot be planes, as an infinitely extended plane would imply an infinite energy flow.

The simplest finite horizon is a static sphere, corresponding to a Schwarzschild black hole. A spherical horizon is characterized by its radius of curvature R, or equivalently, by its surface gravity a; the two quantities are related by $2aR = c^2$. Now, the energy flowing through any horizon is always finite in extension, when measured along the propagation direction. One can thus speak more specifically of an energy pulse. Any energy pulse through a horizon is thus characterized by an energy E and a proper length L. When the energy pulse flows perpendicularly through a horizon, the rate of momentum change, or force, for an observer at the horizon is

$$F = \frac{E}{L} \ . \tag{205}$$

Our goal is to show that the existence of a maximum force implies general relativity. Now, maximum force is realized on horizons. We thus need to insert the maximum possible values on both sides of equation (205) and to show that general relativity follows.

Using the maximum force value and the area $4\pi R^2$ for a spherical horizon we get

$$\frac{c^4}{4G} = \frac{E}{LA} 4\pi R^2 \ . \tag{206}$$

The fraction E/A is the energy per area flowing through any area A that is part of a horizon. The insertion of the maximum values is complete when one notes that the length L of the energy pulse is limited by the radius R. The limit $L \leqslant R$ follows from geometrical considerations: seen from the concave side of the horizon, the pulse must be shorter than the radius of curvature. An independent argument is the following. The length L of
Ref. 289 an object accelerated by a is limited, by special relativity, by $L \leqslant c^2/2a$. Special relativity already shows that this limit is related to the appearance of a horizon. Together with relation (206), the statement that horizons are surfaces of maximum force leads to the following important relation for static, spherical horizons:

$$E = \frac{c^2}{8\pi G} \, a \, A \ . \tag{207}$$

This *horizon equation* relates the energy flow E through an area A of a spherical horizon with surface gravity a. It states that the energy flowing through a horizon is limited, that this energy is proportional to the area of the horizon, and that the energy flow is proportional to the surface gravity. (The horizon equation is also called the *first law of black hole mechanics* or the *first law of horizon mechanics*.) Ref. 290

The above derivation also yields the intermediate result

$$E \leqslant \frac{c^4}{16\pi G} \frac{A}{L} . \qquad (208)$$

This form of the horizon equation states more clearly that no surface other than a horizon can achieve the maximum energy flow, when the area and pulse length (or surface gravity) are given. No other domain of physics makes comparable statements: they are intrinsic to the theory of gravitation.

An alternative derivation of the horizon equation starts with the emphasis on power instead of on force, using $P = E/T$ as the initial equation.

It is important to stress that the horizon equations (207) and (208) follow from only two assumptions: first, there is a maximum speed in nature, and secondly, there is a maximum force (or power) in nature. No specific theory of gravitation is assumed. The horizon equation might even be testable experimentally, as argued below. (We also note that the horizon equation – or, equivalently, the force or power limit – implies a maximum mass change rate in nature given by $\mathrm{d}m/\mathrm{d}t \leqslant c^3/4G$.)

Next, we have to generalize the horizon equation from static and spherical horizons to general horizons. Since the maximum force is assumed to be valid for *all* observers, whether inertial or accelerating, the generalization is straightforward. For a horizon that is irregularly curved or time-varying the horizon equation becomes

$$\delta E = \frac{c^2}{8\pi G} a \, \delta A . \qquad (209)$$

This differential relation – it might be called the *general horizon equation* – is valid for any horizon. It can be applied separately for every piece δA of a dynamic or spatially changing horizon. The general horizon equation (209) has been known to be equivalent to general relativity at least since 1995, when this equivalence was (implicitly) shown by Jacobson. Ref. 291 We will show that the differential horizon equation has the same role for general relativity as the equation $\mathrm{d}x = c \, \mathrm{d}t$ has for special relativity. From now on, when we speak of the horizon equation, we mean the general, differential form (209) of the relation.

It is instructive to restate the behaviour of energy pulses of length L in a way that holds for any surface, even one that is not a horizon. Repeating the above derivation, one gets

$$\frac{\delta E}{\delta A} \leqslant \frac{c^4}{16\pi G} \frac{1}{L} . \qquad (210)$$

Equality is only realized when the surface A is a horizon. In other words, whenever the value $\delta E/\delta A$ in a physical system approaches the right-hand side, a horizon starts to form. This connection will be essential in our discussion of apparent counterexamples to the limit principles.

If one keeps in mind that on a horizon the pulse length L obeys $L \leqslant c^2/2a$, it becomes clear that the general horizon equation is a consequence of the maximum force $c^4/4G$ or the maximum power $c^5/4G$. In addition, the horizon equation takes also into account maximum speed, which is at the origin of the relation $L \leqslant c^2/2a$. The horizon equation thus follows purely from these two limits of nature.

The remaining part of the argument is simply the derivation of general relativity from Ref. 291 the general horizon equation. This derivation was implicitly provided by Jacobson, and the essential steps are given in the following paragraphs. (Jacobson did not stress that his derivation was valid also for continuous space-time, or that his argument could also be used in classical general relativity.) To see the connection between the general horizon equation (209) and the field equations, one only needs to generalize the general horizon equation to general coordinate systems and to general directions of energy–momentum flow. This is achieved by introducing tensor notation that is adapted to curved space-time.

To generalize the general horizon equation, one introduces the general surface element $\mathrm{d}\Sigma$ and the local boost Killing vector field k that generates the horizon (with suitable norm). Jacobson uses these two quantities to rewrite the left-hand side of the general horizon equation (209) as

$$\delta E = \int T_{ab} k^a \mathrm{d}\Sigma^b \ , \qquad (211)$$

where T_{ab} is the energy–momentum tensor. This expression obviously gives the energy at the horizon for arbitrary coordinate systems and arbitrary energy flow directions.

Jacobson's main result is that the factor $a\,\delta A$ in the right hand side of the general horizon equation (209) can be rewritten, making use of the (purely geometric) Raychaudhuri equation, as

$$a\ \delta A = c^2 \int R_{ab} k^a \mathrm{d}\Sigma^b \ , \qquad (212)$$

where R_{ab} is the Ricci tensor describing space-time curvature. This relation describes how the local properties of the horizon depend on the local curvature.

Combining these two steps, the general horizon equation (209) becomes

$$\int T_{ab} k^a \mathrm{d}\Sigma^b = \frac{c^4}{8\pi G} \int R_{ab} k^a \mathrm{d}\Sigma^b \ . \qquad (213)$$

Jacobson then shows that this equation, together with local conservation of energy (i.e., vanishing divergence of the energy–momentum tensor) can only be satisfied if

$$T_{ab} = \frac{c^4}{8\pi G} \left(R_{ab} - (\frac{R}{2} + \Lambda) g_{ab} \right) \ , \qquad (214)$$

where R is the Ricci scalar and Λ is a constant of integration the value of which is not determined by the problem. The above equations are the full field equations of general relativity, including the cosmological constant Λ. The field equations thus follow from the horizon equation. They are therefore shown to be valid at horizons.

Since it is possible, by choosing a suitable coordinate transformation, to position a horizon at any desired space-time point, the field equations must be valid over the whole of space-time. This observation completes Jacobson's argument. Since the field equations follow, via the horizon equation, from the maximum force principle, we have also shown that at every space-time point in nature the same maximum force holds: the value of the maximum force is an invariant and a constant of nature.

In other words, the field equations of general relativity are a direct consequence of the limit on energy flow at horizons, which in turn is due to the existence of a maximum force (or power). In fact, as Jacobson showed, the argument works in both directions. Maximum force (or power), the horizon equation, and general relativity are equivalent.

In short, *the maximum force principle is a simple way to state that, on horizons, energy flow is proportional to area and surface gravity*. This connection makes it possible to deduce the full theory of general relativity. In particular, a maximum force value is sufficient to tell space-time how to curve. We will explore the details of this relation shortly. Note that if no force limit existed in nature, it would be possible to 'pump' any desired amount of energy through a given surface, including any horizon. In this case, the energy flow would not be proportional to area, horizons would not have the properties they have, and general relativity would not hold. We thus get an idea how the maximum flow of energy, the maximum flow of momentum and the maximum flow of mass are all connected to horizons. The connection is most obvious for black holes, where the energy, momentum Page 442 or mass are those falling into the black hole.

By the way, since the derivation of general relativity from the maximum force principle or from the maximum power principle is now established, we can rightly call these limits *horizon force* and *horizon power*. Every experimental or theoretical confirmation of the field equations indirectly confirms their existence.

Space-time is curved

Imagine two observers who start moving parallel to each other and who continue straight ahead. If after a while they discover that they are not moving parallel to each other any Challenge 656 ny more, then they can deduce that they have moved on a curved surface (try it!) or in a curved space. In particular, this happens near a horizon. The derivation above showed that a finite maximum force implies that all horizons are curved; the curvature of horizons in turn implies the curvature of space-time. If nature had only flat horizons, there would be no space-time curvature. The existence of a maximum force implies that space-time is curved.

A horizon so strongly curved that it forms a closed boundary, like the surface of a Page 441 sphere, is called a black hole. We will study black holes in detail below. The main property of a black hole, like that of any horizon, is that it is impossible to detect what is 'behind' the boundary.*

The analogy between special and general relativity can thus be carried further. In special relativity, maximum speed implies $\mathrm{d}x = c\,\mathrm{d}t$, and the change of time depends on the observer. In general relativity, maximum force (or power) implies the horizon equation $\delta E = \frac{c^2}{8\pi G}\, a\, \delta A$ and the observation that space-time is curved.

* Analogously, in special relativity it is impossible to detect what moves faster than the light barrier.

The maximum force (or power) thus has the same double role in general relativity as the maximum speed has in special relativity. In special relativity, the speed of light is the maximum speed; it is also the proportionality constant that connects space and time, as the equation $dx = c\,dt$ makes apparent. In general relativity, the horizon force is the maximum force; it also appears (with a factor 2π) in the field equations as the proportionality constant connecting energy and curvature. The maximum force thus describes both the elasticity of space-time and – if we use the simple image of space-time as a medium – the maximum tension to which space-time can be subjected. This double role of a material constant as proportionality factor and as limit value is well known in materials science.

Does this analogy make you think about aether? Do not worry: physics has no need for the concept of aether, because it is indistinguishable from vacuum. General relativity does describe the vacuum as a sort of material that can be deformed and move. Page 537

Why is the maximum force also the proportionality factor between curvature and energy? Imagine space-time as an elastic material. The elasticity of a material is described by a numerical material constant. The simplest definition of this material constant is the ratio of stress (force per area) to strain (the proportional change of length). An exact definition has to take into account the geometry of the situation. For example, the shear modulus G (or μ) describes how difficult it is to move two parallel surfaces of a material against each other. If the force F is needed to move two parallel surfaces of area A and length l against each other by a distance Δl, one defines the shear modulus G by

$$\frac{F}{A} = G\frac{\Delta l}{l} \; . \qquad (215)$$

The shear modulus for metals and alloys ranges between 25 and 80 GPa. The continuum theory of solids shows that for any crystalline solid without any defect (a 'perfect' solid) there is a so-called theoretical shear stress: when stresses higher than this value are applied, the material breaks. The theoretical shear stress, in other words, the maximum stress in a material, is given by

$$G_{\text{tss}} = \frac{G}{2\pi} \; . \qquad (216)$$

The maximum stress is thus essentially given by the shear modulus. This connection is similar to the one we found for the vacuum. Indeed, imagining the vacuum as a material that can be bent is a helpful way to understand general relativity. We will use it regularly in the following. Ref. 292

What happens when the vacuum is stressed with the maximum force? Is it also torn apart like a solid? Yes: in fact, when vacuum is torn apart, particles appear. We will find out more about this connection later on: since particles are quantum entities, we need to study quantum theory first, before we can describe the effect in the last part of our mountain ascent.

Conditions of validity of the force and power limits

The maximum force value is valid only under certain assumptions. To clarify this point, we can compare it to the maximum speed. The speed of light (in vacuum) is an upper limit for motion of systems with *momentum* or *energy* only. It can, however, be exceeded for motions of non-material points. Indeed, the cutting point of a pair of scissors, a laser light spot on the Moon, or the group velocity or phase velocity of wave groups can exceed the Page 278 speed of light. In addition, the speed of light is a limit only if measured *near* the moving mass or energy: the Moon moves faster than light if one turns around one's axis in a second; distant points in a Friedmann universe move apart from each other with speeds larger than the speed of light. Finally, the observer must be *realistic*: the observer must be made of matter and energy, and thus move more slowly than light, and must be able to Ref. 293 observe the system. No system moving at or above the speed of light can be an observer.

The same three conditions apply in general relativity. In particular, relativistic gravity forbids point-like observers and test masses: they are not realistic. Surfaces moving faster than light are also not realistic. In such cases, counterexamples to the maximum force Challenge 657 ny claim can be found. Try and find one – many are possible, and all are fascinating. We will explore some of the most important ones.

Gedanken experiments and paradoxes about the force limit

> Wenn eine Idee am Horizonte eben aufgeht, ist gewöhnlich die Temperatur der Seele dabei sehr kalt. Erst allmählich entwickelt die Idee ihre Wärme, und am heissesten ist diese (das heisst sie tut ihre grössten Wirkungen), wenn der Glaube an die Idee schon wieder im Sinken ist.
>
> Friedrich Nietzsche*

The last, but central, step in our discussion of the force limit is the same as in the discussion of the speed limit. We need to show that any *imaginable* experiment – not only any real one – satisfies the hypothesis. Following a tradition dating back to the early twentieth century, such an imagined experiment is called a *Gedanken experiment*, from the German Gedankenexperiment, meaning 'thought experiment'.

In order to dismiss all imaginable attempts to exceed the maximum speed, it is sufficient to study the properties of velocity addition and the divergence of kinetic energy near the speed of light. In the case of maximum force, the task is much more involved. Indeed, stating a maximum force, a maximum power and a maximum mass change easily provokes numerous attempts to contradict them. We will now discuss some of these.

▪ The brute force approach. The simplest attempt to exceed the force limit is to try to accelerate an object with a force larger than the maximum value. Now, acceleration implies the transfer of energy. This transfer is limited by the horizon equation (209) or the limit (210). For any attempt to exceed the force limit, the flowing energy results in the appearance of a horizon. But a horizon prevents the force from exceeding the limit, because it imposes a limit on interaction.

Page 309 We can explore this limit directly. In special relativity we found that the acceleration

* 'When an idea is just rising on the horizon, the soul's temperature with respect to it is usually very cold. Only gradually does the idea develop its warmth, and it is hottest (which is to say, exerting its greatest influence) when belief in the idea is already once again in decline.' Friedrich Nietzsche (1844–1900), German philosopher and scholar. This is aphorism 207 – *Sonnenbahn der Idee* – from his text *Menschliches Allzumenschliches – Der Wanderer und sein Schatten*.

of an object is limited by its length. Indeed, at a distance given by $c^2/2a$ in the direction opposite to the acceleration a, a *horizon* appears. In other words, an accelerated body breaks, at the latest, at that point. The force F on a body of mass M and radius R is thus limited by

$$F \leqslant \frac{M}{2R} c^2 . \tag{217}$$

It is straightforward to add the (usually small) effects of gravity. To be observable, an accelerated body must remain *larger* than a black hole; inserting the corresponding radius $R = 2GM/c^2$ we get the force limit (202). *Dynamic* attempts to exceed the force limit thus fail.

▪ The rope attempt. We can also try to generate a higher force in a *static* situation, for example by pulling two ends of a rope in opposite directions. We assume for simplicity that an unbreakable rope exists. To produce a force exceeding the limit value, we need to store large (elastic) energy in the rope. This energy must enter from the ends. When we increase the tension in the rope to higher and higher values, more and more (elastic) energy must be stored in smaller and smaller distances. To exceed the force limit, we would need to add more energy per distance and area than is allowed by the horizon equation. A horizon thus inevitably appears. But there is no way to stretch a rope across a horizon, even if it is unbreakable. A horizon leads either to the breaking of the rope or to its detachment from the pulling system. Horizons thus make it impossible to generate forces larger than the force limit. In fact, the assumption of infinite wire strength is unnecessary: the force limit cannot be exceeded even if the strength of the wire is only finite.

We note that it is not important whether an applied force pulls – as for ropes or wires – or pushes. In the case of *pushing* two objects against each other, an attempt to increase the force value without end will equally lead to the formation of a horizon, due to the limit provided by the horizon equation. By definition, this happens precisely at the force limit. As there is no way to use a horizon to push (or pull) on something, the attempt to achieve a higher force ends once a horizon is formed. Static forces cannot exceed the limit value.

▪ The braking attempt. A force limit provides a maximum momentum change per time. We can thus search for a way to *stop* a moving physical system so abruptly that the maximum force might be exceeded. The non-existence of rigid bodies in nature, already known from special relativity, makes a completely sudden stop impossible; but special relativity on its own provides no lower limit to the stopping time. However, the inclusion of gravity does. Stopping a moving system implies a transfer of energy. The energy flow per area cannot exceed the value given by the horizon equation. Thus one cannot exceed the force limit by stopping an object.

Page 309

Similarly, if a rapid system is *reflected* instead of stopped, a certain amount of energy needs to be transferred and stored for a short time. For example, when a tennis ball is reflected from a large wall its momentum changes and a force is applied. If many such balls are reflected at the same time, surely a force larger than the limit can be realized? It turns out that this is impossible. If one attempted it, the energy flow at the wall would reach the limit given by the horizon equation and thus create a horizon. In that case, no reflection is possible any more. So the limit cannot be exceeded.

▪ The classical radiation attempt. Instead of systems that pull, push, stop or reflect *mat-*

ter, we can explore systems where *radiation* is involved. However, the arguments hold in exactly the same way, whether photons, gravitons or other particles are involved. In particular, mirrors, like walls, are limited in their capabilities.

It is also impossible to create a force larger than the maximum force by concentrating a large amount of light onto a surface. The same situation as for tennis balls arises: when the limit value E/A given by the horizon equation (210) is reached, a horizon appears that prevents the limit from being broken.

▪ The brick attempt. The force and power limits can also be tested with more concrete Gedanken experiments. We can try to exceed the force limit by stacking weight. But even building an infinitely high brick tower does not generate a sufficiently strong force on its foundations: integrating the weight, taking into account its decrease with height, yields a finite value that cannot reach the force limit. If we continually increase the mass density of the bricks, we need to take into account that the tower and the Earth will change into a black hole. And black holes, as mentioned above, do not allow the force limit to be exceeded.

▪ The boost attempt. A boost can apparently be chosen in such a way that a force Ref. 294 value F in one frame is transformed into any desired value F' in another frame. However, this result is not physical. To be more concrete, imagine a massive observer, measuring the value F, at rest with respect to a large mass, and a second observer moving towards the charged mass with relativistic speed, measuring the value F'. Both observers can be thought as being as small as desired. If one transforms the force field at rest F applying the Lorentz transformations, the force F' for the moving observer can reach extremely high values, as long as the speed is high enough. However, a force must be measured by an observer located at the specific point. One has thus to check what happens when the rapid observer moves towards the region where the force is supposed to exceed the force limit. Suppose the observer has a mass m and a radius r. To be an observer, it must be larger than a black hole; in other words, its radius must obey $r > 2Gm/c^2$, implying that the observer has a non-vanishing size. When the observer dives into the force field surrounding the sphere, there will be an energy flow E towards the observer determined by the transformed field value and the crossing area of the observer. This interaction energy can be made as small as desired, by choosing a sufficiently small observer, but the energy is never zero. When the moving observer approaches the large massive charge, the interaction energy increases. Before the observer arrives at the point where the force was supposed to be higher than the force limit, the interaction energy will reach the horizon limits (209) or (210) for the observer. Therefore, a horizon appears and the moving observer is prevented from observing anything at all, in particular any value above the horizon force.

The same limitation appears when electrical or other interactions are studied using a test observer that is charged. In summary, boosts cannot beat the force limit.

▪ The divergence attempt. The force on a test mass m at a radial distance d from a Ref. 288 Schwarzschild black hole (for $\Lambda = 0$) is given by

$$F = \frac{GMm}{d^2\sqrt{1-\frac{2GM}{dc^2}}} \,. \tag{218}$$

In addition, the inverse square law of universal gravitation states that the force between two masses m and M is

$$F = \frac{GMm}{d^2} . \tag{219}$$

Both expressions can take any value; this suggest that no maximum force limit exists.

A detailed investigation shows that the maximum force still holds. Indeed, the force in the two situations diverges only for non-physical point-like masses. However, the maximum force implies a minimum approach distance to a mass m given by

$$d_{\text{min}} = \frac{2Gm}{c^2} . \tag{220}$$

The minimum approach distance – in simple terms, this would be the corresponding black hole radius – makes it impossible to achieve zero distance between two masses or between a horizon and a mass. This implies that a mass can never be point-like, and that there is a (real) minimum approach distance, proportional to the mass. If this minimum approach distance is introduced in equations (218) and (219), one gets

$$F = \frac{c^4}{4G} \frac{Mm}{(M+m)^2} \frac{1}{\sqrt{1 - \frac{M}{M+m}}} \leqslant \frac{c^4}{4G} \tag{221}$$

and

$$F = \frac{c^4}{4G} \frac{Mm}{(M+m)^2} \leqslant \frac{c^4}{4G} . \tag{222}$$

The maximum force value is thus never exceeded, as long as we take into account the size of observers.

- The consistency problem. If observers cannot be point-like, one might question whether it is still correct to apply the original definition of momentum change or energy change as the integral of values measured by observers attached to a given surface. In general relativity, observers cannot be point-like, but they can be as small as desired. The original definition thus remains applicable when taken as a limit procedure for ever-decreasing observer size. Obviously, if quantum theory is taken into account, this limit procedure comes to an end at the Planck length. This is not an issue for general relativity, as long as the typical dimensions in the situation are much larger than this value.
- The quantum problem. If quantum effects are neglected, it is possible to construct surfaces with sharp angles or even fractal shapes that overcome the force limit. However, such surfaces are not physical, as they assume that lengths smaller than the Planck length can be realized or measured. The condition that a surface be physical implies that it must have an intrinsic uncertainty given by the Planck length. A detailed study shows that quantum effects do not allow the horizon force to be exceeded.

Challenge 658 ny

Ref. 286

- The relativistically extreme observer attempt. Any extreme observer, whether in rapid inertial or in accelerated motion, has no chance to beat the limit. In classical physics we are used to thinking that the interaction necessary for a measurement can be made as small as desired. This statement, however, is not valid for all observers; in particular, extreme observers cannot fulfil it. For them, the measurement interaction is large. As a result, a horizon forms that prevents the limit from being exceeded.

▪ The microscopic attempt. We can attempt to exceed the force limit by accelerating a small particle as strongly as possible or by colliding it with other particles. High forces do indeed appear when two high energy particles are smashed against each other. However, if the combined energy of the two particles became high enough to challenge the force limit, a horizon would appear before they could get sufficiently close.

In fact, quantum theory gives exactly the same result. Quantum theory by itself already provides a limit to acceleration. For a particle of mass m it is given by Ref. 295

$$a \leqslant \frac{2mc^3}{\hbar} \; . \qquad (223)$$

Here, $\hbar = 1.1 \cdot 10^{-34}$ Js is the quantum of action, a fundamental constant of nature. In particular, this acceleration limit is satisfied in particle accelerators, in particle collisions and in pair creation. For example, the spontaneous generation of electron–positron pairs in intense electromagnetic fields or near black hole horizons does respect the limit (223). Inserting the maximum possible mass for an elementary particle, namely the (corrected) Planck mass, we find that equation (223) then states that the horizon force is the upper bound for elementary particles. Page 995

▪ The compaction attempt. Are black holes really the most dense form of matter or energy? The study of black hole thermodynamics shows that mass concentrations with higher density than black holes would contradict the principles of thermodynamics. In black hole thermodynamics, surface and entropy are related: reversible processes that reduce entropy could be realized if physical systems could be compressed to smaller values than the black hole radius. As a result, the size of a black hole is the limit size for a mass in nature. Equivalently, the force limit cannot be exceeded in nature. Ref. 288

▪ The force addition attempt. In special relativity, composing velocities by a simple vector addition is not possible. Similarly, in the case of forces such a naive sum is incorrect; any attempt to add forces in this way would generate a horizon. If textbooks on relativity had explored the behaviour of force vectors under addition with the same care with which they explored that of velocity vectors, the force bound would have appeared much earlier in the literature. (Obviously, general relativity is required for a proper treatment.)

Challenge 659 ny
▪ Can you propose and resolve another attempt to exceed the force or power limit?

Gedanken experiments with the power limit and the mass flow limit

Like the force bound, the power bound must be valid for all *imaginable* systems. Here are some attempts to refute it.

▪ The cable-car attempt. Imagine an engine that accelerates a mass with an unbreakable and massless wire (assuming that such a wire could exist). As soon as the engine reached the power bound, either the engine or the exhausts would reach the horizon equation. When a horizon appears, the engine cannot continue to pull the wire, as a wire, even an infinitely strong one, cannot pass a horizon. The power limit thus holds whether the engine is mounted inside the accelerating body or outside, at the end of the wire pulling it.

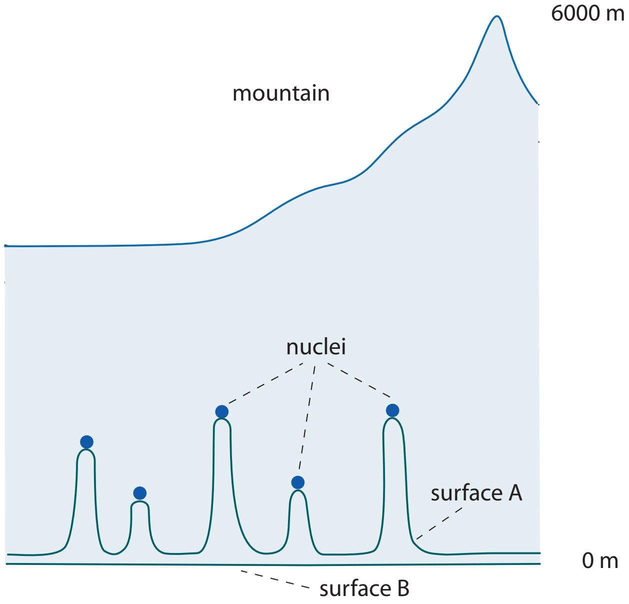


Figure 169 The mountain attempt to exceed the maximum mass flow value

▪ The mountain attempt. It is possible to define a surface that is so strangely bent that it passes *just below* every nucleus of every atom of a mountain, like the surface A in Figure 169. All atoms of the mountain above sea level are then *just above* the surface, barely touching it. In addition, imagine that this surface is moving *upwards* with almost the speed of light. It is not difficult to show that the mass flow through this surface is higher than the mass flow limit. Indeed, the mass flow limit $c^3/4G$ has a value of about 10^{35} kg/s. In a time of 10^{-22} s, the diameter of a nucleus divided by the speed of light, only 10^{13} kg need to flow through the surface: that is the mass of a mountain.

This surface seems to provide a counterexample to the limit. However, a closer look shows that this is not the case. The problem is the expression 'just below'. Nuclei are quantum particles and have an indeterminacy in their position; this indeterminacy is essentially the nucleus–nucleus distance. As a result, in order to be sure that the surface of interest has all atoms *above* it, the shape cannot be that of surface A in Figure 169. It must be a flat plane that remains below the whole mountain, like surface B in the figure. However, a flat surface beneath a mountain does not allow the mass change limit to be exceeded.

▪ The multiple atom attempt. One can imagine a number of atoms equal to the number of the atoms of a mountain that all lie with large spacing (roughly) in a single plane. Again, the plane is moving upwards with the speed of light. But also in this case the uncertainty in the atomic positions makes it impossible to say that the mass flow limit has been exceeded.

▪ The multiple black hole attempt. Black holes are typically large and the uncertainty in their position is thus negligible. The mass limit $c^3/4G$, or power limit $c^5/4G$, corresponds to the flow of a single black hole moving through a plane at the speed of light. Several black holes crossing a plane together at just under the speed of light thus seem to beat the limit. However, the surface has to be physical: an observer must be possible on each of its points. But no observer can cross a black hole. A black hole thus effectively punctures the plane surface. No black hole can ever be said to cross a plane surface; even less so a multiplicity of black holes. The limit remains valid.

- The multiple neutron star attempt. The mass limit seems to be in reach when several neutron stars (which are slightly less dense than a black hole of the same mass) cross a plane surface at the same time, at high speed. However, when the speed approaches the speed of light, the crossing time for points far from the neutron stars and for those that actually cross the stars differ by large amounts. Neutron stars that are almost black holes cannot be crossed in a short time in units of a coordinate clock that is located far from the stars. Again, the limit is not exceeded.
- The luminosity attempt. The existence of a maximum luminosity bound has been Ref. 288 discussed by astrophysicists. In its full generality, the maximum bound on power, i.e. on energy per time, is valid for any energy flow *through any physical surface whatsoever*. The physical surface may even run across the whole universe. However, not even bringing together all lamps, all stars and all galaxies of the universe yields a surface which has a larger power output than the proposed limit.

The surface must be *physical*.* A surface is *physical* if an observer can be placed on each of its points. In particular, a physical surface may not cross a horizon, or have local detail Page 925 finer than a certain minimum length. This minimum length will be introduced later on; it is given by the corrected Planck length. If a surface is not physical, it may provide a Challenge 660 n counterexample to the power or force limits. However, these counterexamples make no statements about nature. (*Ex falso quodlibet.***)

- The many lamp attempt. An absolute power limit imposes a limit on the rate of energy transport through any imaginable surface. At first sight, it may seem that the combined power emitted by two radiation sources that each emit 3/4 of the maximum value should give 3/2 times that value. However, two such lamps would be so massive that they would form a black hole. No amount of radiation that exceeds the limit can leave. Again, since the horizon limit (210) is achieved, a horizon appears that swallows the light and prevents the force or power limit from being exceeded.
- The light concentration attempt. Another approach is to shine a powerful, short and spherical flash of light onto a spherical mass. At first sight it seems that the force and power limits can be exceeded, because light energy can be concentrated into small volumes. However, a high concentration of light energy forms a black hole or induces the mass to form one. There is no way to pump energy into a mass at a faster rate than that dictated by the power limit. In fact, it is impossible to group light sources in such a way that their total output is larger than the power limit. Every time the force limit is approached, a horizon appears that prevents the limit from being exceeded.
- The black hole attempt. One possible system in nature that actually *achieves* the power limit is the final stage of black hole evaporation. However, even in this case the power limit is not exceeded, but only equalled.
- The water flow attempt. One could try to pump water as rapidly as possible through a large tube of cross section A. However, when a tube of length L filled with water flowing

* It can also be called *physically sensible*.

** Anything can be deduced from a falsehood.

at speed v gets near to the mass flow limit, the gravity of the water *waiting* to be pumped through the area A will slow down the water that is being pumped through the area. The limit is again reached when the cross section A turns into a horizon.

Checking that no system – from microscopic to astrophysical – ever exceeds the maximum power or maximum mass flow is a further test of general relativity. It may seem easy to find a counterexample, as the surface may run across the whole universe or envelop any number of elementary particle reactions. However, no such attempt succeeds.

In summary, in all situations where the force, power or mass-flow limit is challenged, whenever the energy flow reaches the black hole mass–energy density in space or the corresponding momentum flow in time, an event horizon appears; this horizon makes it impossible to exceed the limits. All three limits are confirmed both in observation and in theory. Values exceeding the limits can neither be generated nor measured. Gedanken experiments also show that the three bounds are the tightest ones possible. Obviously, all three limits are open to future tests and to further Gedanken experiments. (If you can think of a good one, let me know.)

Challenge 661 ny

Hide and seek

The absence of horizons in everyday life is the *first* reason why the maximum force principle remained undiscovered for so long. Experiments in everyday life do not highlight the force or power limits. The *second* reason why the principle remained hidden is the erroneous belief in point particles. This is a theoretical reason. (Prejudices against the concept of force in general relativity have also been a factor.) The principle of maximum force – or of maximum power – has thus remained hidden for so long because of a 'conspiracy' of nature that hid it both from theorists and from experimentalists.

For a thorough understanding of general relativity it is essential to remember that point particles, point masses and point-like observers do not exist. They are approximations only applicable in Galilean physics or in special relativity. In general relativity, horizons prevent their existence. The habit of believing that the size of a system can be made as small as desired while keeping its mass constant prevents the force or power limit from being noticed.

An intuitive understanding of general relativity

> Wir leben zwar alle unter dem gleichen Himmel, aber wir haben nicht alle den gleichen Horizont.*
> Konrad Adenauer

The concepts of horizon force and horizon power can be used as the basis for a direct, intuitive approach to general relativity.

▪ What is gravity? Of the many possible answers we will encounter, we now have the first: gravity is the 'shadow' of the maximum force. Whenever we experience gravity as weak, we can remember that a different observer at the same point and time would experience the maximum force. Searching for the precise properties of that observer is a good exercise. Another way to put it: if there were no maximum force, gravity would not exist.

* 'We all live under the same sky, but we do not have the same horizon.' Konrad Adenauer, German chancellor.

▪ The maximum force implies universal gravity. To see this, we study a simple planetary system, i.e., one with small velocities and small forces. A simple planetary system of size L consists of a (small) satellite circling a central mass M at a radial distance $R = L/2$. Let a be the acceleration of the object. Small velocity implies the condition $aL \ll c^2$, deduced from special relativity; small force implies $\sqrt{4GMa} \ll c^2$, deduced from the force limit. These conditions are valid for the system as a whole and for all its components. Both expressions have the dimensions of speed squared. Since the system has only one characteristic speed, the two expressions $aL = 2aR$ and $\sqrt{4GMa}$ must be proportional, yielding

$$a = f\frac{GM}{R^2} , \tag{224}$$

where the numerical factor f must still be determined. To determine it, we study the escape velocity necessary to leave the central body. The escape velocity must be smaller than the speed of light for any body larger than a black hole. The escape velocity, derived from expression (224), from a body of mass M and radius R is given by $v_{esc}^2 = 2fGM/R$. The minimum radius R of objects, given by $R = 2GM/c^2$, then implies that $f = 1$. Therefore, for low speeds and low forces, the inverse square law describes the orbit of a satellite around a central mass.

▪ If empty space-time is elastic, like a piece of metal, it must also be able to oscillate. Any physical system can show oscillations when a deformation brings about a restoring force. We saw above that there is such a force in the vacuum: it is called gravitation. In other words, vacuum must be able to oscillate, and since it is extended, it must also be able to sustain waves. Indeed, gravitational waves are predicted by general relativity, as
Page 368 we will see below.

▪ If curvature and energy are linked, the maximum speed must also hold for gravitational energy. Indeed, we will find that gravity has a finite speed of propagation. The inverse square law of everyday life cannot be correct, as it is inconsistent with any speed limit. More about the corrections induced by the maximum speed will become clear shortly. In addition, since gravitational waves are waves of massless energy, we would expect the maximum speed to be their propagation speed. This is indeed the case, as we
Page 368 will see.

▪ A body cannot be denser than a (non-rotating) black hole of the same mass. The maximum force and power limits that apply to horizons make it impossible to squeeze mass into smaller horizons. The maximum force limit can therefore be rewritten as a limit for the size L of physical systems of mass m:

$$L \geqslant \frac{4Gm}{c^2} . \tag{225}$$

If we call twice the radius of a black hole its 'size', we can state that no physical system of mass m is smaller than this value.* The size limit plays an important role in general relativity. The opposite inequality, $m \geqslant \sqrt{A/16\pi}\, c^2/G$, which describes the maximum

* The maximum value for the mass to size limit is obviously equivalent to the maximum mass change given above.

'size' of black holes, is called the *Penrose inequality* and has been proven for many physically realistic situations. The Penrose inequality can be seen to imply the maximum force limit, and vice versa. The maximum force principle, or the equivalent minimum size of matter–energy systems, thus prevents the formation of naked singularities, and implies the validity of the so-called *cosmic censorship*. Ref. 296, Ref. 297, Ref. 298

- There is a power limit for all energy sources. In particular, the value $c^5/4G$ limits the luminosity of all gravitational sources. Indeed, all formulae for gravitational wave emission imply this value as an upper limit. Furthermore, numerical relativity simulations never exceed it: for example, the power emitted during the simulated merger of two black holes is below the limit. Ref. 288
- Perfectly plane waves do not exist in nature. Plane waves are of infinite extension. But neither electrodynamic nor gravitational waves can be infinite, since such waves would carry more momentum per time through a plane surface than is allowed by the force limit. The non-existence of plane gravitational waves also precludes the production of singularities when two such waves collide.
- In nature, there are no infinite forces. There are thus no naked singularities in nature. Horizons prevent the appearance of naked singularities. In particular, the big bang was *not* a singularity. The mathematical theorems by Penrose and Hawking that seem to imply the existence of singularities tacitly assume the existence of point masses – often in the form of 'dust' – in contrast to what general relativity implies. Careful re-evaluation of each such proof is necessary.
- The force limit means that space-time has a limited stability. The limit suggests that space-time can be torn into pieces. This is indeed the case. However, the way that this happens is not described general relativity. We will study it in the third part of this text.
- The maximum force is the standard of force. This implies that the gravitational constant G is constant in space and time – or at least, that its variations across space and time cannot be detected. Present data support this claim to a high degree of precision. Ref. 299
- The maximum force principle implies that gravitational energy – as long as it can be defined – *falls* in gravitational fields in the same way as other type of energy. As a result, the maximum force principle predicts that the Nordtvedt effect vanishes. The Nordtvedt effect is a hypothetical periodical change in the orbit of the Moon that would appear if the gravitational energy of the Earth–Moon system did not fall, like other mass–energy, in the gravitational field of the Sun. Lunar range measurements have confirmed the absence of this effect. Ref. 288
- If horizons are surfaces, we can ask what their colour is. This question will be explored later on. Page 441
- Later on we will find that quantum effects cannot be used to exceed the force or power limit. Quantum theory also provides a limit to motion, namely a lower limit to action; however, this limit is independent of the force or power limit. (A dimensional analysis already shows this: there is no way to define an action by combinations of c and G.) Therefore, even the combination of quantum theory and general relativity does not help in overcoming the force or power limits.

An intuitive understanding of cosmology

A maximum power is the simplest possible explanation of Olbers' paradox. Power and luminosity are two names for the same observable. The sum of all luminosities in the universe is finite; the light and all other energy emitted by all stars, taken together, is finite. If one assumes that the universe is homogeneous and isotropic, the power limit $P \leqslant c^5/4G$ must be valid across any plane that divides the universe into two halves. The part of the universe's luminosity that arrives on Earth is then so small that the sky is dark at night. In fact, the actually measured luminosity is still smaller than this calculation, as a large part of the power is not visible to the human eye (since most of it is matter anyway). In other words, the night is dark because of nature's power limit. This explanation is *not* in contrast to the usual one, which uses the finite lifetime of stars, their finite density, their finite size, and the finite age and the expansion of the universe. In fact, the combination of all these usual arguments simply implies and repeats in more complex words that the power limit cannot be exceeded. However, this more simple explanation seems to be absent in the literature.

The existence of a maximum force in nature, together with homogeneity and isotropy, implies that the visible universe is of *finite size*. The opposite case would be an infinitely large, homogeneous and isotropic universe. But in that case, any two halves of the universe would attract each other with a force above the limit (provided the universe were sufficiently old). This result can be made quantitative by imagining a sphere whose centre lies at the Earth, which encompasses all the universe, and whose radius decreases with time almost as rapidly as the speed of light. The mass flow $\mathrm{d}m/\mathrm{d}t = \rho A v$ is predicted to reach the mass flow limit $c^3/4G$; thus one has

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \rho_0 4\pi R_0^2 c = \frac{c^3}{4G} \quad , \tag{226}$$

Ref. 300 a relation also predicted by the Friedmann models. The precision measurements of the cosmic background radiation by the WMAP satellite confirm that the present-day total energy density ρ_0 (including dark matter and dark energy) and the horizon radius R_0 just reach the limit value. The maximum force limit thus predicts the observed size of the universe.

A finite power limit also suggests that a finite age for the universe can be deduced. Can you find an argument? Challenge 662 ny

Experimental challenges for the third millennium

The lack of direct tests of the horizon force, power or mass flow is obviously due to the lack of horizons in the environment of all experiments performed so far. Despite the difficulties in reaching the limits, their values are observable and falsifiable.

In fact, the force limit might be tested with high-precision measurements in binary pulsars or binary black holes. Such systems allow precise determination of the positions of the two stars. The maximum force principle implies a relation between the position error Δx and the energy error ΔE. For all systems one has Ref. 286

$$\frac{\Delta E}{\Delta x} \leqslant \frac{c^4}{4G} \; . \tag{227}$$

For example, a position error of 1 mm gives a mass error of below $3 \cdot 10^{23}$ kg. In everyday life, all measurements comply with this relation. Indeed, the left side is so much smaller than the right side that the relation is rarely mentioned. For a direct check, only systems which might achieve direct equality are interesting. Dual black holes or dual pulsars are such systems.

It might be that one day the amount of matter falling into some black hole, such as the one at the centre of the Milky Way, might be measured. The limit $\mathrm{d}m/\mathrm{d}t \leqslant c^3/4G$ could then be tested directly.

The power limit implies that the highest luminosities are only achieved when systems emit energy at the speed of light. Indeed, the maximum emitted power is only achieved when all matter is radiated away as rapidly as possible: the emitted power $P = Mc^2/(R/v)$ cannot reach the maximum value if the body radius R is larger than that of a black hole (the densest body of a given mass) or the emission speed v is lower than that of light. The sources with highest luminosity must therefore be of maximum density and emit entities without rest mass, such as gravitational waves, electromagnetic waves or (maybe) gluons. Candidates to detect the limit are black holes in formation, in evaporation or undergoing mergers.

A candidate surface that reaches the limit is the night sky. The night sky is a horizon. Provided that light, neutrino, particle and gravitational wave flows are added together, the limit $c^5/4G$ is predicted to be reached. If the measured power is smaller than the limit (as it seems to be at present), this might even give a hint about new particles yet to be discovered. If the limit were exceeded or not reached, general relativity would be shown to be incorrect. This might be an interesting future experimental test.

The power limit implies that a wave whose integrated intensity approaches the force limit cannot be plane. The power limit thus implies a limit on the product of intensity I (given as energy per unit time and unit area) and the size (curvature radius) R of the front of a wave moving with the speed of light c:

$$4\pi R^2 I \leqslant \frac{c^5}{4G} \; . \tag{228}$$

Obviously, this statement is difficult to check experimentally, whatever the frequency and type of wave might be, because the value appearing on the right-hand side is extremely large. Possibly, future experiments with gravitational wave detectors, X-ray detectors, gamma ray detectors, radio receivers or particle detectors might allow us to test relation (228) with precision. (You might want to predict which of these experiments will confirm the limit first.)

Challenge 663 e

The lack of direct experimental tests of the force and power limits implies that *indirect tests* become particularly important. All such tests study the motion of matter or energy and compare it with a famous consequence of the force and power limits: the field equations of general relativity. This will be our next topic.

A summary of general relativity

There is a simple axiomatic formulation of general relativity: the horizon force $c^4/4G$ and the horizon power $c^5/4G$ are the highest possible force and power values. No contradicting observation is known. No counterexample has been imagined. General relativity follows from these limits. Moreover, the limits imply the darkness of the night and the finiteness of the size of the universe.

The principle of maximum force has obvious applications for the teaching of general relativity. The principle brings general relativity to the level of first-year university, and possibly to well-prepared secondary school, students: only the concepts of maximum force and horizon are necessary. space-time curvature is a consequence of horizon curvature.

Challenge 664 ny

Page 424

The concept of a maximum force points to an additional aspect of gravitation. The cosmological constant Λ is not fixed by the maximum force principle. (However, the principle does fix its sign to be positive.) Present measurements give the result $\Lambda \approx 10^{-52}/\mathrm{m}^2$. A positive cosmological constant implies the existence of a negative energy volume density $-\Lambda c^4/G$. This value corresponds to a negative pressure, as pressure and energy density have the same dimensions. Multiplication by the (numerically corrected) Planck area $2G\hbar/c^3$, the smallest area in nature, gives a force value

Page 993

$$F = 2\Lambda\hbar c = 0.60 \cdot 10^{-77}\,\mathrm{N}\,. \qquad (229)$$

This is also the gravitational force between two (numerically corrected) Planck masses $\sqrt{\hbar c/8G}$ located at the cosmological distance $1/4\sqrt{\Lambda}$. If we make the somewhat wishful assumption that expression (229) is the smallest possible force in nature (the numerical factors are not yet verified), we get the fascinating conjecture that the full theory of general relativity, including the cosmological constant, may be defined by the combination of a *maximum* and a *minimum* force in nature. (Can you find a smaller force?)

Challenge 665 ny

Proving the minimum force conjecture is more involved than for the case of the maximum force. So far, only some hints are possible. Like the maximum force, the minimum force must be compatible with gravitation, must not be contradicted by any experiment, and must withstand any Gedanken experiment. A quick check shows that the minimum force, as we have just argued, allows us to deduce gravitation, is an invariant, and is not contradicted by any experiment. There are also hints that there may be no way to generate or measure a smaller value. For example, the minimum force corresponds to the energy per length contained by a photon with a wavelength of the size of the universe. It is hard – but maybe not impossible – to imagine the production of a still smaller force.

We have seen that the maximum force principle and general relativity fail to fix the value of the cosmological constant. Only a unified theory can do so. We thus get two requirements for such a theory. First, any unified theory must predict the same upper limit to force. Secondly, a unified theory must fix the cosmological constant. The appearance of $\hbar$ in the conjectured expression for the minimum force suggests that the minimum force is determined by a combination of general relativity and quantum theory. The proof of this suggestion and the direct measurement of the minimum force are two important challenges for our ascent beyond general relativity.

We are now ready to explore the consequences of general relativity and its field equations in more detail. We start by focusing on the concept of space-time curvature in everyday life, and in particular, on its consequences for the observation of motion.

Bibliography

284 The author thanks Steve Carlip, Corrado Massa, Tom Helmond and Peter Brown for interesting discussions. Cited on page 319.

285 It seems that the first published statement of the principle was in an earlier edition of this text, in the chapter on gravitation and relativity. The reference is C. Schiller, *Motion Mountain – A Hike Through and Beyond Space and Time Following the Concepts of Modern Physics*, found at http://www.motionmountain.net. The approach was discussed in various usenet discussion groups in the early twenty-first century. Cited on page 319.

286 C. Schiller, Maximum force and minimum distance: physics in limit statements, part of this text and downloadable at http://www.motionmountain.net/C19-LIMI.pdf. Cited on pages 319, 321, 323, 332, and 339.

287 G.W. Gibbons, The maximum tension principle in general relativity, *Foundations of Physics* **32**, pp. 1891–1901, 2002, or http://www.arxiv.org/abs/hep-th/0210109. Cited on page 319.

288 H.C. Ohanian & R. Ruffini, *Gravitation and Spacetime*, W.W. Norton & Co., New York, 1994. Cited on pages 323, 331, 333, 335, 338, 394, and 396.

289 See for example Wolfgang Rindler, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001, p. 70 ff, or Ray d'Inverno *Introducing Einstein's Relativity*, Clarendon Press, 1992, p. 36 ff. Cited on page 324.

290 See for example A. Ashtekar, S. Fairhust & B. Krishnan, Isolated horizons: Hamiltonian evolution and the first law, http://www.arxiv.org/abs/gr-qc/0005083. Cited on page 325.

291 T. Jacobson, Thermodynamics of spacetime: the Einstein equation of state, *Physical Review Letters* **75**, pp. 1260–1263, 1995 or http://www.arxiv.org/abs/gr-qc/9504004. Cited on pages 325 and 326.

292 See for example Ekkehart Kröner, *Kontinuumstheorie der Versetzungen und Eigenspannungen*, Springer, 1958, volume 5 of the series 'Ergebnisse der angewandten Mathematik'. Kröner shows the similarity between the equations, methods and results of solid-state continuum physics and those of general relativity. Cited on page 328.

293 Edwin F. Taylor & John A. Wheeler, *Spacetime Physics – Introduction to Special Relativity*, second edition, Freeman, 1992. Cited on page 329.

294 This counterexample was suggested by Steve Carlip. Cited on page 331.

295 E.R. Caianiello, *Lettere al Nuovo Cimento* **41**, p. 370, 1984. Cited on page 333.

296 R. Penrose, Naked singularities, *Annals of the New York Academy of Sciences* **224**, pp. 125–134, 1973. Cited on page 338.

297 G. Huisken & T. Ilmanen, The Riemannian Penrose inequality, *Int. Math. Res. Not.* **59**, pp. 1045–1058, 1997. Cited on page 338.

298 S.A. Hayward, Inequalities relating area, energy, surface gravity and charge of black holes, *Physical Review Letters* **81**, pp. 4557–4559, 1998. Cited on page 338.

299 C. Will, *Was Einstein Right? – Putting General Relativity to the Test*, Oxford University Press, 1993. See also his paper http://www.arxiv.org/abs/gr-qc/9811036. Cited on page 338.

300 The measurements by the WMAP satellite are summarized in ... Cited on page 339.

8. The new ideas on space, time and gravity

Sapere aude.

Horace*

Gravitational influences do transport energy.** In the description of motion, the next oal must therefore be to increase the precision in such a way that this transport happens at most with the speed of light. Henri Poincaré stated this requirement already in 1905. The results following from this principle will be fascinating; we will find that empty space can move, that the universe has a finite age and that objects can be in permanent free fall. It will turn out that empty space can be bent but that it is much stiffer than steel. Despite these strange statements, the theory and all its predictions have been confirmed by each one of the numerous experiments ever performed.

Describing motion due to gravity using the relation $a = GM/r^2$ allows speeds larger than light. Indeed, the speed of a mass in orbit is not limited. It is also unclear how the values of a and r depend on the observer. Universal gravitation thus cannot be correct. In order to reach the correct description, called *general relativity* by Albert Einstein, we
Ref. 301, Ref. 302 have to throw quite a few preconceptions overboard.

Rest and free fall

The opposite of motion in daily life is a body at rest, such as a child sleeping. Or a rock defying the waves. A body is at rest whenever it is not disturbed by other bodies. In the Galilean description of the world, rest is the *absence of velocity*. In special relativity, rest became *inertial motion*, since no inertially moving observer can distinguish its own motion from rest: nothing disturbs him. Both the rock in the waves and the rapid protons crossing the galaxy as cosmic rays are at rest. Including gravity leads us to an even more general definition.

Challenge 666 e If any body moving inertially is to be considered at rest, then any body in free fall must also be. Nobody knows this better than Joseph Kittinger, the man who in August
Ref. 303 1960 stepped out of a balloon capsule at the record height of 31.3 km. At that altitude, the air is so thin that during the first minute of his free fall he felt completely at rest, as if he were floating. Being an experienced parachutist, he was so surprised that he had to turn upwards in order to convince himself that he was indeed getting away from his balloon! Despite his lack of any sensation, he was falling at up to 274 m/s or 988 km/h with respect to the Earth's surface. He only started feeling something from the moment in which he encountered the first substantial layers of air. That was when his free fall started to be disturbed. Later, after four and a half minutes of fall, his special parachute opened and nine minutes later he landed in New Mexico.

Kittinger and all other observers in free fall, such as the cosmonauts circling the Earth, make the same observation: it is impossible to distinguish anything happening in free fall from what would happen at rest. This impossibility is called the *principle of equivalence*; it is one of the starting points of general relativity. It leads to the most precise – and final – definition of rest: *rest is free fall*. Rest is lack of disturbance; so is free fall.

* 'Venture to be wise.' Quintus Horatius Flaccus, Ep. 1, 2, 40.
** The details of this statement are far from simple. They are discussed on page 368 and page 398.

The set of all free falling observers that meet at a point in space-time generalize the set of the inertial observers that can meet at a point in special relativity. This means that we must describe motion in such a way that not only inertial but also freely falling observers can talk to each other. In addition, a full description of motion must be able to describe gravitation and the motion it produces, and it must be able to describe motion for any observer imaginable. This is the aim that general relativity realizes.

To pursue this aim, we put the result in simple words: true *motion is the opposite of free fall.* This conclusion directly produces a number of questions: Most trees or mountains are not in free fall, thus they are not at rest. What motion are they undergoing? And if free fall is rest, what is weight? And what then is gravity anyway? Let us start with the last question.

Challenge 667 n

What is gravity? – A second answer

In the beginning, we defined gravity as the shadow of the maximum force. But there is a second way to describe it, more related to everyday life. As William Unruh likes to explain, the constancy of the speed of light for all observers implies a simple conclusion: *gravity is the uneven running of clocks at different places.** Of course, this seemingly absurd definition needs to be checked. The definition does not talk about a single situation seen by *different* observers, as we often did in special relativity. The definition states that neighbouring, identical clocks, fixed against each other, run differently in the presence of a gravitational field when watched by the *same* observer; moreover, this difference is defined to be what we usually call gravity. There are two ways to check this connection: by experiment and by reasoning. Let us start with the latter method, as it is cheaper, faster and more fun.

Ref. 304
Challenge 669 e

An observer feels no difference between gravity and constant acceleration. We can thus study constant acceleration and use a way of reasoning we encountered already in the chapter on special relativity. We assume light is emitted at the back end of a train of length Δh accelerating forward with acceleration g. The light arrives at the front after a time $t = \Delta h/c$. However, during this time the accelerating train has picked up some additional velocity, namely $\Delta v = gt = g\Delta h/c$. As a result, due to the Doppler effect we encountered in special relativity, the frequency f of the light arriving at the front has changed. Inserting, we get**

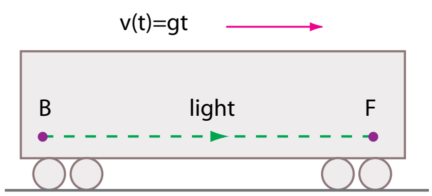


Figure 170 Inside an accelerating train or bus

Page 277
Challenge 670 e

$$\frac{\Delta f}{f} = \frac{g\Delta h}{c^2} . \qquad (230)$$

The sign of the frequency change depends on whether the light motion and the train

* Gravity is also the uneven length of meter bars at different places, as we will see below. Both effects are needed to describe it completely; but for daily life on Earth, the clock effect is sufficient, since it is much larger than the length effect, which can usually be neglected. Can you see why?

Challenge 668 n

** The expression $v = gt$ is valid only for non-relativistic speeds; nevertheless, the conclusion of the section is independent of this approximation.

Challenge 671 e

acceleration are in the same or in opposite directions. For actual trains or buses, the frequency change is quite small; nevertheless, it does appear. Acceleration induces frequency changes in light. Let us compare this effect of acceleration with the effects of gravity.

Challenge 672 n

To measure time and space, we use light. What happens to light when gravity is involved? The simplest experiment is to let light fall or rise. In order to deduce what must happen, we add a few details. Imagine a conveyor belt carrying masses around two wheels, a low and a high one. The descending, grey masses are slightly larger, as shown in Figure 171. Whenever such a larger mass is near the bottom, some mechanism – not drawn – converts the mass surplus to light via $E = mc^2$ and sends the light up towards the top.* At the top, one of the lighter, white masses passing by absorbs the light and, due to its added weight, turns the conveyor belt until it reaches the bottom. Then the process repeats.**

Ref. 305

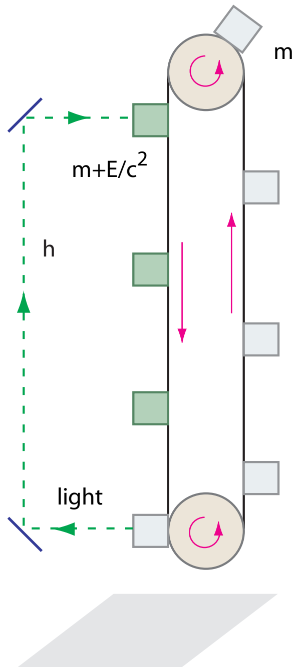


Figure 171 The necessity of blue- and red-shift of light: why trees are greener at the bottom

As the grey masses on the descending side are always heavier, the belt would turn for ever and this system could continuously *generate* energy. However, since energy conservation is at the basis of our definition of time, as we saw in the beginning of our walk, the whole process must be impossible. We have to conclude that the light changes its energy when climbing. The only possibility is that the light arrives at the top with frequency *different* from the one at which it is emitted from the bottom.***

Page 182

In short, it turns out that rising light is gravitationally red-shifted. Similarly, the light descending from the top of a tree down to an observer is *blue shifted*; this gives a darker, older colour to the top in comparison to the bottom of the tree. General relativity thus says that trees have different shades of green along their height.**** How big is the effect? The result deduced from the drawing is again the one of formula (230). That is expected, as light moving in an accelerating train and light moving in gravity are equivalent situations, as you might want to check yourself. The formula gives a relative change of frequency f of only $1.1 \cdot 10^{-16}$ /m on the surface of the Earth. For trees, this so-called *gravitational red-shift* or *gravitational Doppler effect* is far too small to be observable, at least using normal light.

Challenge 676 e
Challenge 677 n

In 1911, Einstein proposed to check the change of frequency with height by measuring the red-shift of light emitted by the Sun, using the famous Fraunhofer lines as colour markers. The first experiments, by Schwarzschild and others, were unclear or even negative, due to a number

Ref. 306
Page 703

* As in special relativity, here and in the rest of our mountain ascent, the term 'mass' always refers to rest mass. Note that the conversion is always lossless.

Challenge 673 e

** Can this process be performed with 100% efficiency?

Challenge 674 n

*** The precise relation between energy and frequency of light is described and explained in the part on quantum theory, on page 670. But we know already from classical electrodynamics that the energy of light depends on its intensity and on its frequency.

**** How does this argument change if you include the illumination by the Sun?

Challenge 675 ny

of other effects that induce colour changes at high temperatures. After some unclear first measurements, in 1920 and 1921, Grebe and Bachem, and independently Perot, confirmed the gravitational red-shift with careful experiments. In later years, technology advances made the measurements much easier, until it was even possible to measure the effect on Earth. In 1960, in a classic experiment using the Mössbauer effect, Pound and Rebka confirmed the gravitational red-shift in their university tower using γ radiation. Ref. 307

But our two Gedanken experiments tell us much more. Let us use the same arguments as in the case of special relativity: a colour change implies that clocks run *differently* at the top and at the bottom, as they do in the front and in the back of a train. The time difference $\Delta\tau$ is predicted to depend on the height difference Δh and the acceleration of gravity g like

$$\frac{\Delta\tau}{\tau} = \frac{\Delta f}{f} = \frac{g\Delta h}{c^2} \; . \tag{231}$$

Therefore, in gravity, *time is height dependent*. That was exactly what we claimed above. In fact, *height makes old*. Can you confirm this conclusion? Challenge 678 ny

In 1972, by flying four precise clocks in an aeroplane while keeping an identical one on the ground, Hafele and Keating found that clocks indeed run differently at different altitudes according to expression (231). Subsequently, in 1976, the team of Vessot *et al.* shot a precision clock based on a maser – a precise microwave generator and oscillator – upwards on a missile. The team compared the maser inside the missile with an identical maser on the ground and again confirmed the expression. In 1977, Briatore and Leschiutta showed that a clock in Torino indeed ticks more slowly than one on the top of the Monte Rosa. They confirmed the prediction that on Earth, for every 100 m of height gained, people age more rapidly by about 1 ns per day. In the mean time this effect has been confirmed for all systems for which experiments were performed, such as several other planets, the Sun and numerous other stars. Ref. 308 Ref. 309 Ref. 310 Challenge 679 ny

Do these experiments show that time changes or are they simply due to clocks that function badly? Take some time and try to settle the question. One argument only is given: gravity does change the colour of light, and thus really does change time. Clock precision is not an issue here. Challenge 680 e

In summary, gravity is indeed the uneven running of clocks at different heights. Note that both an observer at the lower position and one at a higher position *agree* on the result; both find that the upper clock goes faster. In other words, when gravity is present, space-time is *not* described by the Minkowski space-time of special relativity, but by some more general space-time. To put it mathematically, whenever gravity is present, the 4-distance $\mathrm{d}s^2$ between events is different from the expression without gravity:

$$\mathrm{d}s^2 \neq c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \; . \tag{232}$$

We will give the correct expression shortly.

Is this view of gravity as height-dependent time really reasonable? No. It turns out that it is not yet strange enough. Since the speed of light is the same for all observers, we can say more. If time changes with height, also length must do so! More precisely, if clocks run differently at different heights, also the length of meter bars changes with height. Can you confirm this for the case of horizontal bars at different heights? Challenge 681 ny

If length changes with height, the circumference of a circle around the Earth *cannot* be given by $2\pi r$. A similar discrepancy is also found by an ant measuring radius and circumference of a large circle traced on the surface of a basketball. Indeed, gravity implies that humans are in a similar situation to ants on a basketball, with the only difference that the circumstances are translated from two to three dimensions. We conclude that wherever gravity plays a role, space is *curved*.

What tides tell us about gravity

During his free fall, Kittinger was able to specify an inertial frame for himself. Indeed, he felt completely at rest. Does this mean that it is impossible to distinguish acceleration from gravitation? No; distinction *is* possible. We only have to compare *two* (or more) falling observers.

Kittinger could not have found a frame which is also inertial for a colleague falling on the opposite side of the Earth. Such a common frame does not exist. In general, it is impossible to find a *single* inertial reference frame describing different observers freely falling near a mass. In fact, the impossibility to find a common inertial frame applies even to *nearby* observers in a gravitational field. Two nearby observers observe that during fall, their relative distance changes. (Why?) The same happens to orbiting observers.

Challenge 682 e

Challenge 683 n

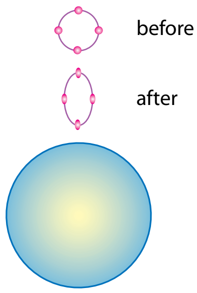


Figure 172 Tidal effects: what bodies feel when falling

In a closed room in orbit around the Earth, a person or a mass at the centre of the room would not feel anything, and in particular no gravity. But if several particles are located in the room, they will behave differently depending on their exact position in the room. Only if two particles were on exactly the same orbit they would keep their relative position. If one particle is in a lower or higher orbit than the other, they will depart from each other over time. Even more interestingly, if a particle in orbit is displaced sideways, it will oscillate around the central position. (Can you confirm this?)

Challenge 684 ny

Gravitation leads to relative distance change. The distance changes evince another effect, shown in Figure 172: an extended body in free fall is slightly *squeezed*. Also this effect tells us that the essence of gravity is that free fall is *different* from point to point. That rings a bell. The squeezing of the body is the same effect that leads to the tides. Indeed, the bulging oceans can be seen as the squeezed Earth in its fall towards the Moon. Using this result of universal gravity we can now affirm: the essence of gravity is the observation of tidal effects.

Page 120
Ref. 311

In other words, gravity is simple only *locally*. Only locally does it look like acceleration. Only locally, a falling observer like Kittinger feels at rest. In fact, only a point-like observer does so! As soon as we take spatial extension into account, we find tidal effects. *Gravity is the presence of tidal effects.* The absence of tidal effects implies the absence of gravity. Tidal effects are the everyday consequence of height-dependent time. Isn't this a beautiful conclusion?

In principle, Kittinger could have *felt* gravitation during his free fall, even with his eyes closed, had he paid attention to himself. Had he measured the distance change between his two hands, he would have found a tiny decrease which could have told him that he was falling, even with his eyes closed. This tiny decrease would have forced Kittinger to a strange conclusion. Two inertially moving hands should move along two parallel lines, always keeping the same distance. Since the distance changes, in the space around him lines starting out in parallel do not remain so. Kittinger would have concluded that the space around him was similar to the surface of a sphere, where two lines starting out north, parallel to each other, also change distance, until they meet at the North Pole. In other words, Kittinger would have concluded that he was in a *curved* space.

Studying the value of the distance decrease between his hands, Kittinger would even have concluded that the curvature of space changes with height. Physical space differs from a sphere, which has constant curvature; physical space is more involved. The effect is extremely small and cannot be felt by human senses. Kittinger had no chance to detect anything. Detection requires special high sensitivity apparatus. However, the conclusion does not change. Space-time is *not* described by Minkowski space when gravity is present. Tidal effects imply space-time curvature. Gravity is curved space-time.

Bent space and mattresses

> Wenn ein Käfer über die Oberfläche einer Kugel krabbelt, merkt er wahrscheinlich nicht, daß der Weg, den er zurücklegt, gekrümmt ist. Ich dagegen hatte das Glück, es zu merken.*
>
> Albert Einstein's answer to his son Eduard's question about the reason for his fame

On the 7th of November 1919, Albert Einstein became world famous. On that day, the Times newspaper in London announced the results of a double expedition to South America with the title 'Revolution in science / new theory of the universe / Newtonian ideas overthrown'. The expedition had shown unequivocally – though not for the first time – that the theory of universal gravity, essentially given by $a = GM/r^2$, was wrong, and that instead space had been shown to be *curved*. A worldwide mania started. Einstein was presented as the greatest of all geniuses. 'Space warped' was the most common headline. Einstein's papers on general relativity were reprinted in full in popular magazines. People could read the field equations of general relativity, in tensor form and with Greek indices, in the middle of Time magazine. This did not happen to any other physicist before or afterwards. What was the reason for this excitement?

The expedition to the southern hemisphere had performed an experiment proposed
Ref. 312 by Einstein himself. Apart from searching for the change of time with height, Einstein had also thought about a number of experiments to detect the curvature of space. In the one that eventually made him famous, Einstein proposed to take a picture of the stars near the Sun, as is possible during a solar eclipse, and compare it with a picture of the same stars at night, when the Sun is far away. Einstein predicted a change in position of 1.75' (1.75 seconds of arc) for star images at the border of the Sun, a result *twice* as large as
Page 123 the effect predicted by universal gravity. The prediction, corresponding to about 1/40 mm on the photographs, was confirmed in 1919, and thus universal gravity was ruled out.

* When a bug walks over the surface of a sphere it probably does not notice that the path it walks is curved. I, on the other hand, had the luck to notice it.

Does this experiment *imply* that space is curved? Not by itself. In fact, other explanations could be given for the result of the eclipse experiment, such as a potential differing from the one of universal gravity. However, the eclipse results are not the only data. We already know about the change of time with height. Experiments show that any two observers at different height measure the same value for the speed of light c near themselves. But these experiments also show that if an observer measures the speed of light at the position of the *other* observer, he gets a value *differing* from c, since his clock runs differently. There is only one possible solution to this dilemma: meter bars, like clocks, also *change* with height, and in such a way as to yield the same speed of light everywhere.

If the speed of light is constant but clocks and meter bars change with height, *space is curved near masses*. Many physicists in the twentieth century checked whether meter bars really behave differently in places where gravity is present. And indeed, curvature has been detected around several planets, around all the hundreds of stars where it could be measured, and around dozens of galaxies. Many indirect effects of curvature around masses, to be described in detail below, have also been observed. All results confirm the existence of curvature of space and space-time around masses, and in addition confirm the predicted curvature values. In other words, meter bars near masses do indeed change their size from place to place, and even from orientation to orientation. Figure 173 gives an impression of the situation.

Challenge 685 ny

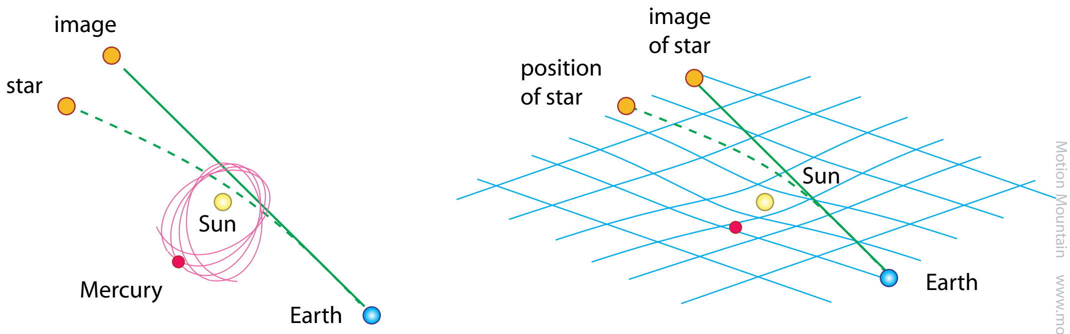


Figure 173 The mattress model of space: the path of a light beam and of a satellite near a spherical mass

But attention: the right hand figure, even though found in all textbooks, can be misleading. It can be easily mistaken to show a *potential* around a body. Indeed, it is impossible to draw a graph showing curvature and potential separately. (Why?) We will see that for small curvatures, it is even possible to describe the meter bar change with a potential only. Thus the figure does not really cheat, at least in the case of weak gravity. But for large and changing values of gravity, a potential cannot be defined, and thus there is indeed no way to avoid using curved space to describe gravity. In summary, if we imagine space as a sort of generalized mattress in which masses produce deformations, we have a reasonable model of space-time. As masses move, the deformation follows them.

Ref. 313
Challenge 686 n

The mattress also shows that the acceleration of a test particle does only depend on the curvature of the mattress. It does not depend on the mass of the test particles. In other words, the mattress model explains why all bodies fall in the same way. (In the old days, this was also called the equality of the inertial and gravitational mass.)

Space thus behaves like a frictionless mattress that pervades everything. We live inside the mattress, but we do not feel it in everyday life. Massive objects pull the foam of the mattress towards them, thus deforming the shape of the mattress. More force, more energy or more mass imply a larger deformation than smaller values. (Does the mattress Page 537 make you think about aether? Do not worry; physics eliminated the concept of aether because aether and vacuum are indistinguishable. This is exactly what we do here.)

If gravity means curved space, we deduce that any accelerated observer, such as a man in a departing car, must also observe that space is curved. However, in everyday life we do not notice any such effect, because for accelerations and sizes of usual accelerated observers the resulting curvature values are too small to be noticed. Could you devise a Challenge 687 ny precision experiment to check the prediction?

Curved space-time

Even though Figure 173 shows the curvature of space only, not only space, but also space-time is curved. We will shortly find out how to describe both the shape of space as well as the shape of space-time, and how to measure their curvature.

Let us have a first idea on how to describe nature when space-time is curved. In the case of Figure 173, the best description of events is the use of the time t shown by a clock located at spatial infinity; that avoids problems with the uneven running of clocks with different distances from the central mass. For the radial coordinate r the most practical choice to avoid problems with the curvature of space is to use the circumference of a circle around the central body divided by 2π. The curved shape of space-time is best described Page 267 by the behaviour of the space-time distance $\mathrm{d}s$, or by the wristwatch time $\mathrm{d}\tau = \mathrm{d}s/c$, between two neighbouring points with coordinates (t, r) and $(t + \mathrm{d}t, r + \mathrm{d}r)$. We found Page 347 out above that gravity means that in spherical coordinates we have

$$\mathrm{d}\tau^2 = \frac{\mathrm{d}s^2}{c^2} \neq \mathrm{d}t^2 - \mathrm{d}r^2/c^2 - r^2\mathrm{d}\varphi^2/c^2 \ . \tag{233}$$

The inequality expresses that space-time is *curved*. Indeed, the experiments on time change with height show that the space-time interval around a spherical mass is given by

$$\mathrm{d}\tau^2 = \frac{\mathrm{d}s^2}{c^2} = \left(1 - \frac{2GM}{rc^2}\right)\mathrm{d}t^2 - \frac{\mathrm{d}r^2}{c^2 - \frac{2GM}{r}} - \frac{r^2}{c^2}\mathrm{d}\varphi^2 \ . \tag{234}$$

This expression is called the *Schwarzschild metric* after one of its discoverers.* The metric (234) describes the curved shape of space-time around a spherical non-rotating mass. It is Challenge 688 n well approximated by the Earth or the Sun. (Why can the rotation be neglected?) Gravity's

* Karl Schwarzschild (1873–1916), important German astronomer; he was one of the first persons to understand general relativity. He published his solution in December 1915, only a few months after Einstein had published his field equations. He died prematurely, at age 42, much to Einstein's distress. We will deduce the metric later on, directly from the field equations of general relativity. The other discoverer of the metric, Ref. 314 unknown to Einstein, was the Dutch physicist J. Droste.

strength is obviously measured by a dimensionless number h defined as

$$h = \frac{2G}{c^2}\frac{M}{R} . \tag{235}$$

This ratio expresses the gravitational strain with which lengths and the vacuum are deformed from the flat situation of special relativity, and thus also determines the amount that clocks slow down when gravity is present. (The number also tells how far off one is from any possible horizon.) On the surface of the Earth the ratio h has the small value of $1.4 \cdot 10^{-9}$; on the surface of the Sun is has the somewhat larger value of $4.2 \cdot 10^{-6}$. The precision of modern clocks allow one to detect such small effects quite easily. The various consequences and uses of the deformation of space-time will be discussed shortly.

We note that if a mass is highly concentrated, in particular when its radius gets *equal* to its so-called *Schwarzschild radius*

$$R_S = \frac{2GM}{c^2} , \tag{236}$$

the Schwarzschild metric behaves strangely: at that location, *time disappears* (note that t is time at infinity). At the Schwarzschild radius, the wristwatch time (as shown by a clock at infinity) stops – and a horizon appears. What happens precisely will be explored

Page 443

below. The situation is not common; the Schwarzschild radius for a mass like the Earth is 8.8 mm and for the Sun 3.0 km; you might want to check that the object size for all

Challenge 689 e

systems in everyday life is always larger than their Schwarzschild radius. Bodies which

Ref. 315

reach this limit are called *black holes*; we will study them in detail shortly. In fact, general relativity states that *no* system in nature is smaller than its Schwarzschild size, or that the ratio h defined by expression (235) is never above unity.

In summary, the results mentioned so far make it clear that *mass generates curvature*. Special relativity then tells us that as a consequence, space should also be curved by the presence of any type of energy–momentum. Every type of energy curves space-time. For example, light should also curve space-time. Unfortunately, even the highest energy beams correspond to extremely small masses, and thus to unmeasurably small curvatures. Even heat curves space-time. In most systems, heat is only about a fraction of 10^{-12} of total mass; its curvature effect is thus unmeasurable and negligible. Nevertheless it is still possible to show experimentally that energy also curves space. In almost all atoms a sizeable fraction of the mass is due to the electrostatic energy among the positively charged protons. In 1968 Kreuzer confirmed that energy curves space with a clever experiment using

Ref. 316

a floating mass.

It is straightforward to picture that the uneven running of clock is the temporal equi-

Challenge 690 ny

valent of spatial curvature. Taking the two together, we conclude that the complete statement is that in case of gravity, *space-time* is curved.

Let us sum up our chain of thoughts. Energy is equivalent to mass; mass produces gravity; gravity is equivalent to acceleration; acceleration is position-dependent time. Since light speed is constant, we deduce that *energy–momentum tells space-time to curve*. This statement is the first half of general relativity.

We will soon find out how to measure curvature, how to calculate it from energy–

momentum and how to compare the two results. We will also find out that different observers measure different curvature values. The set of transformations relating one viewpoint to another in general relativity, *diffeomorphism symmetry*, will tell us how to relate the results.

Since matter moves, we can say even more. Not only is space-time curved near masses, it also bends back when a mass has passed by. In other words, general relativity states that space, as well as space-time, is *elastic*. However, it is rather stiff, and quite a lot stiffer than steel.* In fact, to curve a piece of space by 1 % requires an energy density enormously larger than to curve a simple train rail by 1 %. This and other fun consequences of space-time and its elasticity will occupy us for this chapter.

Challenge 691 ny

The speed of light and the constant of gravitation

Si morior, moror.**

We continue on the way towards precision in gravitation. All the experiments and knowledge about gravity can be summed up in just two general statements. The first principle states:

▷ *The speed v of a physical system is bound by the limit*

$$v \leqslant c \tag{237}$$

for all observers, where c is the speed of light.

The description following from this first principle, *special* relativity, is extended to *general* relativity by adding a second principle, characterizing gravitation. There are several possibilities.

▷ *For all observers, the force on a system is limited by*

$$F \leqslant \frac{c^4}{4G} , \tag{238}$$

where G is the universal constant of gravitation.

In short, there is a maximum force in nature. Gravitation leads to attraction of masses. However, this attraction force is limited. An equivalent limit is:

Challenge 692 e

▷ *For all observers, the size L of a system of mass M is limited by*

$$\frac{L}{M} \geqslant \frac{4G}{c^2} . \tag{239}$$

* A good book in popular style on the topic is David Blair & Geoff McNamara, Ripples on a cosmic sea, Allen & Unwin, 1997.

** 'If I rest, I die.' This is the motto of the bird of paradise.

In other words, there is nothing more concentrated in nature than a non-rotating black hole of the same mass. Another way to express the principle of gravitation is the following:

▷ *For all systems, the emitted power is limited by*

$$P \leqslant \frac{c^5}{4G} . \qquad (240)$$

In short, there is a maximum power in nature. The three limits are all equivalent to each other; no exception is known or possible. The limits reduce to the usual definition of gravity in the non-relativistic case. The limits tell us *what* gravity is, namely curvature, and *how* exactly it behaves. The limits allow us to determine the curvature in all situations, Page 323 at all space-time events. As we have seen above, the speed limit together with any of the second principles imply all of general relativity.*

For example, are you able to show that the formula describing gravitational red-shift Challenge 693 ny complies with the general limit (239) on length to mass ratios?

We note that any formula that contains the speed of light c is based on special relativity, and if it contains the constant of gravitation G, it relates to universal gravity. If a formula contains *both* c and G, it is a statement of general relativity. The present chapter frequently underlines this connection.

The mountain ascent so far has taught us that a precise description of motion requires the listing of all allowed viewpoints, their characteristics, their differences, as well as the specification of the transformations from any viewpoint to any other. From now on, *all* viewpoints are allowed, without exception; anybody must be able to talk to anybody else. It makes no difference whether an observer feels gravity, is in free fall, is accelerated or is in inertial motion. Also people who exchange left and right, people who exchange up and down or people who say that the Sun turns around the Earth must be able to talk to each other. This gives a much larger set of viewpoint transformations than in the case of special relativity; it makes general relativity both difficult and fascinating. And since all viewpoints are allowed, the resulting description of motion is *complete*.**

Why does a stone thrown into the air fall back to Earth? – Geodesics

A genius is somebody who makes all possible mistakes in the shortest possible time.

In special relativity, we saw that inertial or free floating motion is that motion which connects two events that requires the *longest* proper time. In the absence of gravity, the motion fulfilling this requirement is *straight* (rectilinear) motion. On the other hand, we Page 54 are also used to think of straightness as the shape of light rays. Indeed, we all are accustomed to check the straightness of an edge by looking along it. Whenever we draw the axes of a physical coordinate system, we imagine either drawing paths of light rays or drawing the motion of freely moving bodies.

* This didactic approach is unconventional. It is possible that is has been pioneered by the present author. Ref. 317 Also Gary Gibbons developed it independently. Earlier references are not known.

** Were it not for a small deviation called quantum theory.

In the absence of gravity, object paths and light paths coincide. However, in the presence of gravity, objects do not move along light paths, as every thrown stone shows. Light does not define spatial straightness any more. In the presence of gravity, both light and matter paths are bent, though by *different* amounts. But the original statement remains: even when gravity is present, bodies follow paths of longest possible proper time. For matter, such paths are called *timelike geodesics*. For light, the paths are called *lightlike* or *null geodesics*.

In other words, *stones fall because they follow geodesics.* Let us perform a few checks of this statement. We note that in space-time, geodesics are the curves with *maximal* length. This is in contrast with the case of pure space, such as the surface of a sphere, where geodesics are the curves of *minimal* length.

▪ Since stones move by maximizing proper time for inertial observers, they also must do so for freely falling observers, as Kittinger would argue. Then they do so for all observers.

▪ If fall is seen as a consequence of the Earth's surface approaching – as we will argue

Challenge 694 ny

later on – we can deduce directly that fall implies a proper time as long as possible. Free fall is motion along geodesics.

▪ We saw above that gravitation follows from the existence of a maximum force. The result can be visualized also in another way. If the gravitational attraction between a central body and a satellite were *stronger* than it is, black holes would be smaller than they are; in that case the maximum force limit and the maximum speed could be exceeded by getting close to such a black hole. If on the other hand, gravitation were *weaker* than it is, there would be observers for which the two bodies would not interact, thus for which they would not form a physical system. In summary, a maximum force of $c^4/4G$ implies universal gravity. There is no difference in stating that all bodies attract through gravitation or in stating that there is a maximum force with the value $c^4/4G$.

▪ Let us turn to an experimental check. If fall is a consequence of curvature, then the path of *all* stones thrown or falling near the Earth must have the *same* curvature in space-time. Take a stone thrown horizontally, a stone thrown vertically, a stone thrown rapidly, or a stone thrown slowly: it takes only two lines to show that *in space-time* all their paths are approximated to

Challenge 695 ny

high precision by circle segments, as shown in Figure 174. All paths have the *same* curvature radius r, given by

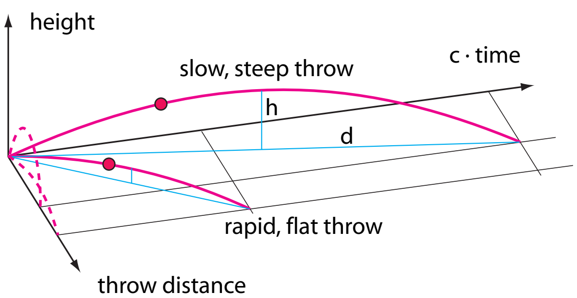


Figure 174 All paths of flying stones have the same curvature in space-time

$$r = \frac{c^2}{g} \approx 9.2 \cdot 10^{15}\ \text{m} . \tag{241}$$

The large value of the radius, corresponding to an extremely low curvature, explains why we do not notice it in everyday life. The parabolic shape typical of the path of a stone in everyday life is just the projection of the more fundamental path in 4-dimensional space-time into 3-dimensional space. The important point is that the value of the curvature does *not* depend on the details of the throw. In fact, this simple result could have brought people onto the path of general relativity already a full century before Einstein; what was missing was the recognition of the importance of the speed of light as limit speed. In any case, this simple calculation confirms that fall and curvature are connected. As expected and mentioned already above, the curvature diminishes at larger heights, until it vanishes at infinite distance from the Earth.

▪ If fall is seen as consequence of the curvature of space-time, the explanation is a bit more involved.

– CS – to be filled in – CS –

In short, the straightest path in space-time for a stone thrown in the air and the path for the thrower himself cross again after a while: stones do fall back. Only if the velocity of the stone is too large, the stone does not fall back: it then leaves the attraction of the Earth and makes its way through the sky.

In summary, the motion of any particle falling freely 'in a gravitational field' is described by the same variational principle as the motion of a free particle in special relativity: the path maximizes the proper time $\int d\tau$. We rephrase this by saying that any particle in free fall from point A to point B minimizes the action S given by

$$S = -mc^2 \int_A^B d\tau \ . \tag{242}$$

That is all we need to know about the free fall of objects. As a consequence, any *deviation from free fall keeps you young*. The larger the deviation, the younger you stay.

Page 461 As we will see below, the minimum action description of free fall has been tested ex-
Ref. 318 tremely precisely, and no difference from experiment has ever been observed. We will also find out that for free fall, the predictions of general relativity and of universal gravity differ substantially both for particles near the speed of light and for central bodies of high density. So far, all experiments showed that whenever the two predictions differ, general relativity is right and universal gravity or other alternative descriptions are wrong.

All bodies fall along geodesics. This connection tells us something important. The fall of bodies does not depend on their mass. The geodesics are like 'rails' in space-time that tell bodies how to fall. In other words, space-time can indeed be imagined like a single, giant, deformed entity. Space-time is an entity of our thinking. The shape of this entity tells objects how to move. Space-time is thus indeed like an intangible mattress; this deformed mattress guides falling objects along its networks of geodesics.

Also *bound* energy falls in the same way as mass, as is proven by comparing the fall of objects made of different materials. They have different percentages of bound energy.
Challenge 696 n (Why?) For example, on the Moon, where there is no air, cosmonauts dropped steel balls and feathers and found that they fell together, alongside each other. The independence
Ref. 319 on material composition has been checked over and over again, and no difference has ever been found.

Can light fall?

How does radiation fall? Light, like any radiation, is energy without rest mass. It moves like a stream of extremely fast and light objects. Therefore deviations from universal gravity become most apparent for light. How does light fall? Light cannot change speed. When
Page 345
light falls vertically, it only changes colour, as we have seen above. But light can also change direction. Already long before relativity, in 1801, the Prussian astronomer Johann
Ref. 320
Soldner understood that universal gravity implies that light is *deflected* when passing near
Page 123
a mass. He also calculated how the deflection angle depends on the mass of the body and the distance of passage. However, nobody in the nineteenth century was able to check the result experimentally.

Obviously, light has energy, and energy has weight; the deflection of light by itself is thus *not* a proof of the curvature of space. Also general relativity predicts a deflection angle for light passing masses, but of *twice* the classical Soldner value, because the curvature of space around large masses adds to the effect of universal gravity. The deflection of light thus only confirms the curvature of space if the *value* agrees with the one predicted by general relativity. This is the case; the observations coincide with the
Page 375
prediction. More details will be given shortly.

Mass is thus not necessary to feel gravity; energy is sufficient. This result of the mass-energy equivalence must become a standard reflex when studying general relativity. In particular, light is not light-weight, but heavy. Can you argue that the curvature of light
Challenge 697 ny
near the Earth must be the same as the one of stones, given by expression (241)?

In summary, all experiments show that not only mass, but also energy falls along geodesics, whatever its type, bound or free, and whatever the interaction, be it electromagnetic or nuclear. Moreover, the motion of radiation confirms that space-time is curved.

Since experiments show that all particles fall in the same way, independently of their mass, charge or any other property, we can conclude that all possible trajectories form an independent structure. This structure is what we call *space-time*.

We thus find that *space-time tells matter, energy and radiation how to fall*. This statement is the second half of general relativity. It complements the first half, which states that energy tells space-time how to curve. To complete the description of macroscopic motion, we only need to add numbers to these statements, so that they become testable. As usual, we can proceed in two ways: we can deduce the equations of motion directly, or we can first deduce the Lagrangian and then deduce the equations of motion from it. But before we do that, let's have some fun.

Curiosities about gravitation

> Wenn Sie die Antwort nicht gar zu ernst nehmen und sie nur als eine Art Spaß ansehen, so kann ich Ihnen das so erklären: Früher hat man geglaubt, wenn alle Dinge aus der Welt verschwinden, so bleiben noch Raum und Zeit übrig. Nach der Relativitätstheorie verschwinden aber auch Zeit und Raum mit den Dingen.*
>
> Albert Einstein in 1921 in New York

General relativity is a beautiful topic with numerous interesting aspects.

▪ Take a plastic bottle and make some holes into it near the bottom. Fill the bottle with water, closing the holes with your fingers. If you let the bottle go, no water will leave the bottle during the fall. Can you explain how this experiment confirms the equivalence of rest and free fall? Challenge 698 ny

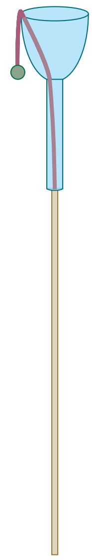


Figure 175 A puzzle

▪ On his 76th birthday, Einstein received a birthday present especially made for him, shown in Figure 175. A rather deep cup is mounted on the top of a broom stick. The cup contains a weak piece of elastic rubber attached to its bottom, to which a ball is attached at the other end. In the starting position, the ball hangs outside the cup. The rubber is too weak to pull the ball into the cup against gravity. What is the most elegant method to get the ball into the cup? Challenge 699 n

▪ The radius of curvature of space-time at the Earth's surface is $9.2 \cdot 10^{15}$ m. Are you able to confirm this value? Challenge 700 e

▪ A piece of wood floats on water. Does it stick out more or less in an elevator accelerating upwards? Challenge 701 ny

▪ We saw in special relativity that if two twins are identically accelerated in the same direction, with one twin some distance ahead of the other, then the twin ahead ages more than the twin behind. Does this happen in a gravitational field as well? And what happens when the field varies with height, as happens on Earth? Page 277 Challenge 702 ny

▪ A maximum force and a maximum power also imply a maximum flow of mass. Can you show that no mass flow can exceed the value $1.1 \cdot 10^{35}$ kg/s? Challenge 703 ny

▪ The experiments of Figure 170 and 171 differ in one point: one happens in flat space, the other in curved space. One seems to be connected with energy conservation, the other not. Do these differences invalidate the arguments given? Challenge 704 ny

▪ How do cosmonauts weigh themselves to check whether they eat enough? Challenge 705 n

▪ Is a cosmonaut really floating freely? No. It turns out that space stations and satellites are accelerated by several small effects. The important ones are the pressure of the light from the Sun, the friction of the thin air, and the effects of solar wind; micrometeorites can usually be neglected. The three effects all lead to accelerations of the order of 10^{-6} m/s^2 to 10^{-8} m/s^2, depending on the height of the orbit. Can you estimate when an apple floating in space will hit the wall of a space station? By the way, what is the magnitude of the tidal accelerations in this situation? Challenge 706 n

▪ There is no negative mass in nature, as discussed in the beginning of our walk (even Page 76

* 'If you do not take the answer too seriously and regard it only for amusement, I can explain it to you in the following way: in the past it was thought that if all things disappear from the world, space and time would remain. But following relativity theory, space and time disappear together with the things.'

antimatter has *positive* mass). This means that gravitation cannot be shielded, in contrast to electromagnetic interactions. Even antimatter has positive mass. Since gravitation cannot be shielded, there is no way to make a perfectly isolated system. But such systems Page 809 form the basis of thermodynamics! We will study the fascinating implications later on; for example, an *upper limit* for the entropy of physical systems will appear.

▪ Can curved space be used to travel faster than light? Imagine a space-time in which two points could be connected either by a path leading through a flat portion of space-time, or by a second path leading through a partially curved portion. Could that curved portion be used to travel between the points faster than through the flat one? Mathematically, this is possible; however, such a curved space would need to have a *negative* energy density. Such a situation is in contrast with the definition of energy and with the Ref. 322 non-existence of negative mass. The statement that this does not happen in nature is also called the *weak energy condition*. Can you say whether it is included in the limit on length Challenge 707 ny to mass ratios?

▪ Like in special relativity, the length to mass limit $L/M \geqslant 4G/c^2$ is a challenge to devise experiments to overcome it. Can you explain what happens when an observer Challenge 708 ny moves so rapidly past a mass that the body's length contraction reaches the limit?

▪ There is an important mathematical aspect which singles out three dimensional space from all other possibilities. A closed (one-dimensional) curve can be knotted *only* in $\mathbf{R}^3$, whereas it can be unknotted in any higher dimension. (The existence of knots is also the reason that three is the smallest dimension that allows chaotic particle motion.) However, general relativity does not say *why* space-time has three plus one dimensions. It is simply based on the fact. This deep and difficult question will be settled only in the third part of the mountain ascent.

▪ Henri Poincaré, who died in 1912, shortly before the general theory of relativity was finished, thought for a while that curved space was not a necessity, but only a possibility. He imagined that one could simply continue using Euclidean space and just add that light Challenge 709 ny follows curved paths. Can you explain why this project is impossible?

▪ Can two hydrogen atoms circle each other, in their respective gravitational field? Challenge 710 n What would the size of this 'molecule' be?

Challenge 711 n ▪ Can two light pulses circle each other, in their respective gravitational field?

▪ The various motions of the Earth mentioned in the section on Galilean physics, such Page 81 as the rotation around its axis or the rotation around the Sun, lead to various types of time in physics and astronomy. The time defined by the best atomic clocks is called *terrestrial dynamical time*. By inserting leap seconds every now and then to compensate for the bad Page 1067 definition of the second (an Earth rotation does not take 86 400, but 86 400.002 seconds) and, in minor ways, for the slowing of Earth's rotation, one gets the *universal time coordinate*. Then there is the time derived from this one by taking into account all leap seconds. One then has the – different – time which would be shown by a non-rotating clock in the centre of the Earth. Finally, there is *barycentric dynamical time*, which is the time that Ref. 323 would be shown by a clock in the centre of mass of the solar system. Only using this latter time can satellites be reliably steered through the solar system. In summary, relativity says goodbye to Greenwich mean time, as does British law, in one of the rare cases were the law follows science.

▪ Space agencies thus *have* to use general relativity if they want to get artificial satellites to Mars, Venus, or comets. Without its use, orbits would not be calculated correctly, and satellites would miss the aimed spots and usually even the whole planet. In fact, space agencies take the safe side; they use a generalization of general relativity, namely the so-called *parametrized post-Newtonian formalism*, which includes a continuous check on whether general relativity is correct. Within measurement errors, no deviation was found so far.*

Ref. 324 ▪ General relativity is also used by space agencies around the world to know the exact positions of satellites and to tune radios to the frequency of radio emitters on them. In addition, general relativity is essential for the so-called *global positioning system*, or GPS. This modern navigation tool* consists of 24 satellites with clocks flying around the world. Why does the system need general relativity to operate? Since both the satellites as well as any person on the surface of the Earth travel in circles, we have $dr = 0$ and we can rewrite the Schwarzschild metric (234) as

$$\left(\frac{\mathrm{d}i}{\mathrm{d}t}\right)^2 = 1 - \frac{2GM}{rc^2} - \frac{r^2}{c^2}\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^2 = 1 - \frac{2GM}{rc^2} - \frac{v^2}{c^2} \,. \tag{244}$$

Challenge 712 e For the relation between satellite time and Earth time we then get

$$\left(\frac{dt_{\mathrm{sat}}}{dt_{\mathrm{Earth}}}\right)^2 = \frac{1 - \frac{2GM}{r_{\mathrm{sat}}c^2} - \frac{v_{\mathrm{sat}}^2}{c^2}}{1 - \frac{2GM}{r_{\mathrm{Earth}}c^2} - \frac{v_{\mathrm{Earth}}^2}{c^2}} \,. \tag{245}$$

Challenge 713 n Can you deduce how many microseconds a satellite clock runs fast every day, given that the GPS satellites turn around the Earth every twelve hours? Since only three microseconds would give a position error of one kilometre after a single day, the clocks in the satellites are adjusted to run slow by the calculated amount. The necessary adjustments are monitored and confirm general relativity every single day within experimental errors, since the system began operation. Ref. 325

Ref. 326 ▪ The gravitational constant G does not seem to change with time. Present experiments limit its rate of change to less than 1 part in 10^{12} per year. Can you imagine how this can be checked? Challenge 714 d

* To give an idea of what this means, the *unparametrized* post-Newtonian formalism, based on general relativity, writes the equation of motion of a body of mass m near a large mass M as a deviation from the inverse square expression for the acceleration a:

$$a = \frac{GM}{r^2} + f_2\frac{GM}{r^2}\frac{v^2}{c^2} + f_4\frac{GM}{r^2}\frac{v^4}{c^4} + f_5\frac{Gm}{r^2}\frac{v^5}{c^5} + \cdots \tag{243}$$

Here the numerical factors f_n are calculated from general relativity and are of order one. The first uneven terms are missing because of the reversibility of general relativistic motion, were it not for gravity wave emission, which accounts for the small term f_5; note that it contains the small mass m instead of the large mass M. Nowadays, all factors f_n up to f_7 have been calculated. However, in the solar system, only the term up to f_2 has ever been detected, a situation which might change with future high precision satellite experiments. Higher order effects, up to f_5, have been measured in the binary pulsars, as discussed below. Page 373

For a *parametrized* post-Newtonian formalism, all factors f_n, including the uneven ones, are fitted through the data coming in; so far all these fits agree with the values predicted by general relativity.

* For more information, see the http://www.gpsworld.com web site.

▪ Could our impression that we live in 3 space dimensions be due to a limitation of
Challenge 715 n
our senses? How?

Challenge 716 ny
▪ Can you estimate the effect of the tides on the colour of the light emitted by an atom?

▪ The strongest possible gravitational field is the one of small black holes. The strongest
Ref. 327
observed gravitational field is somewhat smaller though. In 1998, Zhang and Lamb used the x-ray data from a double star system to determine that space-time near the 10 km sized neutron star is curved up to 30 % of the maximum possible value. What is the corresponding gravitational acceleration, assuming the neutron star has the same mass as
Challenge 717 ny
the Sun?

Ref. 328
▪ Light deflection changes the angular size δ of a mass M with radius r when observed
Challenge 718 e
at distance d. The effect leads to the pretty expression

$$\delta = \arcsin\left(\frac{r\sqrt{1-R_S/d}}{d\sqrt{1-R_S/r}}\right) \quad \text{where} \quad R_S = \frac{2GM}{c^2} . \tag{246}$$

Challenge 719 ny
What is the percentage of the surface of the Sun an observer at infinity can see? We will
Page 450
come back to the issue in more detail shortly.

What is weight?

There is no way that a *single* (and point-like) observer can distinguish the effects of gravity from those of acceleration. This property of nature allows one to make a strange statement: things *fall* because the surface of the Earth accelerates towards them. Therefore, the *weight* of an object results from the surface of the Earth accelerating upwards and pushing against the object. That is the principle of equivalence applied to everyday life. For the same reason, objects in free fall have no weight.

Let us check the numbers. Obviously, an accelerating surface of the Earth produces a weight for each body resting on it. The weight is proportional to the inertial mass. In other words, the inertial mass of a body is exactly identical to the gravitational mass.
Ref. 329
This is indeed observed, and to the highest precision achievable. Roland von Eőtvős* performed many such high-precision experiments throughout his life, without finding any discrepancy. In these experiments, he used the fact that the inertial mass determines centrifugal effects and the gravitational mass determines free fall. Can you imagine how
Challenge 720 ny
exactly he tested the equality?

Page 73
However, the mass equality is not a surprise. Remembering the definition of mass ratio as negative inverse acceleration ratio, independently of the origin of the acceleration, we are reminded that mass measurements cannot be used to distinguish between inertial and gravitational mass at all. We saw that both masses are equal by definition already in
Page 124
Galilean physics, and that the whole discussion is a red herring.

The equality of acceleration and gravity allows us to tell the following story. Imagine stepping into an elevator in order to move down a few stories. Push the button. The following happens: the elevator is pushed upwards by the accelerating surface of the Earth

* Roland von Eőtvős (b. 1848 Budapest, d. 1919 Budapest), Hungarian physicist. He performed many precision gravity experiments; among others, he discovered the effect named after him. The university of Budapest is named after him.

somewhat less than the building; the building overtakes the elevator, which therefore remains behind. Moreover, due to the weaker push, at the beginning everybody inside the elevator feels a bit lighter. When the contact with the building is restored, the elevator is accelerated to catch up with the accelerating surface of the Earth. Therefore we all feel like in a strongly accelerating car, pushed into direction opposite to the acceleration: for a short while, we feel heavier, until the elevator arrives at his destination.

Why do apples fall?

Vires acquirit eundo.

Vergilius*

Sitting in an accelerating car, an object thrown forward will soon be caught by the car again. For the same reason, a stone thrown upwards is soon caught up by the surface of the Earth, which is continuously accelerating upwards. If you enjoy this way of seeing things, imagine an apple falling from a tree. In the moment it detaches, it stops being accelerated upwards by the branch. The apple can now enjoy the calmness of real rest. Our limited human perception calls this state of rest free fall. Unfortunately, the accelerating surface of the Earth approaches mercilessly and, depending on the time the apple stayed at rest, the Earth hits it with a corresponding velocity, leading to more or less severe shape deformation.

Falling apples also teach us not to be disturbed any more by the statement that gravity is the uneven running of clocks with height. In fact, this statement is *equivalent* to saying that the surface of the Earth is accelerating upwards, as the discussion above showed.

Can this reasoning be continued without limit? We can go on for quite a while; it is fun to show how the Earth can be of constant radius even though its surface is accelerating upwards everywhere. We can thus play with the equivalence of acceleration and gravity for quite some time. However, this equivalence is only useful in situations where only one accelerating body is studied. The equivalence between acceleration and gravity ends as soon as *two* falling objects are studied. Any study of several bodies inevitably leads to the conclusion that gravity is not acceleration; *gravity is curved space-time*.

Challenge 721 ny

Many aspects of gravity and curvature can be understood with no or only a little mathematics. The next section will highlight some of the differences between universal gravity and general relativity, showing that only the latter description agrees with experiment. After that, a few concepts for the measurement of curvature are introduced and applied to the motion of objects and space-time. As soon as the reasoning gets too involved for a first reading, skip that section. In any case, the section on the stars, cosmology and black holes again uses only little mathematics.

9. Motion in general relativity – bent light and wobbling vacuum

I have the impression that Einstein understands relativity theory very well.

Chaim Weitzmann, first president of Israel

Before we tackle the details of general relativity, we explore how the motion of objects

* 'Going it acquires strength.' Publius Vergilius Maro (b. 70 Andes, d. 19 BCE Brundisium), *Aeneis* 4, 175.

and light *differs* from that predicted in universal gravity, and how these differences can be measured.

Weak fields

Gravity is strong near horizons. This happens whenever the involved mass M and the involved distance scale R obeys

$$\frac{2GM}{Rc^2} \approx 1 . \tag{247}$$

In other words, gravity is strong mainly in three situations: near black holes, at the horizon of the universe, or at extremely high particle energies. The first two cases are explored later on, the last will be explored in the third part of our mountain ascent. In contrast, in many parts of nature there are *no* nearby horizons; in these cases, gravity is a *weak* effect. Despite the violence of avalanches or of falling asteroids, in everyday life, gravity is much weaker than the maximum force. On the Earth the ratio just mentioned is only about 10^{-9}. In this and all cases of everyday life, gravitation can still be approximated by a field, despite what was said above. These weak field situations are interesting because they are simple to understand; they mainly require for their explanation the different running of clocks with height. Weak field situations allow us to mention space-time curvature only in passing and allow us to continue thinking about gravity as a source of acceleration. Indeed, the change of time with height already induces many new and interesting effects. The only thing we need is a consistent relativistic treatment.

The Thirring effects

In 1918, the Austrian physicist Hans Thirring published two simple and beautiful predictions of motions, one of them with his collaborator Josef Lense. Both motions do not appear in universal gravity, but they do appear in general relativity. Figure 176 shows the predictions. Ref. 330

In the first example, nowadays called the *Thirring effect*, centrifugal accelerations as well as Coriolis accelerations for masses in the interior of a rotating mass shell are predicted. Thirring showed that if an enclosing mass shell rotates, masses inside it are attracted towards the shell. The effect is very small; however, this prediction is in full contrast to universal gravity, where a spherical mass shell – rotating or not – has no effect on masses in their interior. Are you able to explain this effect using the figure and the mattress analogy?

Challenge 722 ny

The second effect, the *Lense–Thirring effect* is more famous. General relativity predicts that an oscillating Foucault pendulum or a satellite circling the Earth in a polar orbit do not stay precisely in a fixed plane compared to the rest of the universe, but that the rotation of the Earth drags the plane along a tiny bit. This *frame-dragging*, as the effect is also called, appears because the Earth in vacuum behaves like a rotating ball in a foamy mattress. When a ball or a shell rotates inside the foam, it partly drags the foam along with it. Similarly, the Earth drags some vacuum with it, and thus turns the plane of the pendulum. For the same reason, the effect also turns the plane of an orbiting satellite.

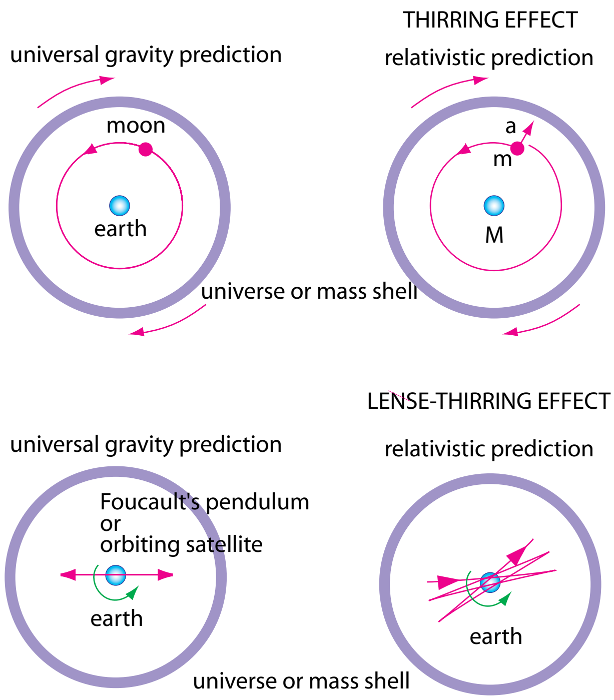


Figure 176 The Thirring and the Lense–Thirring effects

Figure 177 The lageos satellites: metal spheres with a diameter of 60 cm, a mass of 407 kg and covered with 426 retroreflectors

The Lense–Thirring or frame-dragging effect is extremely small. It has been measured for the first time in 1998 by an Italian group led by Ignazio Ciufolini, and then again by the same group in the years up to 2004. They followed the motion of two special artificial satellites – shown in Figure 177 – consisting only of a body of steel and some cat eyes. The group measured the satellite's motion around the Earth with extremely high precision, making use of reflected laser pulses. This method allowed this low budget experiment to beat by many years the efforts of much larger but much more sluggish groups.* The results confirm the tiny predictions by general relativity with an error of about 25 %.

Ref. 331

Frame dragging effects have also been measured in binary star systems. This is possible if one of the stars is a pulsar; such stars send out regular radio pulses, e.g. every millisecond, with extremely high precision. By measuring the exact time when the pulses arrive on Earth, one can deduce the way these stars move and confirm that such subtle effects as frame dragging do take place.

Ref. 332

* One is the so-called Gravity Probe B satellite experiment, which should significantly increase the measurement precision; the satellite has been put in orbit in 2005, after 30 years of planning.

Gravitomagnetism*

Frame-dragging and the Lense–Thirring effect can be seen as special cases of gravitomagnetism. (We will show the connection below.) This approach to gravity, already studied Ref. 333 in the nineteenth century by Holzmüller and by Tisserand, has become popular again in recent years, especially for its didactic aspect. As mentioned above, talking about a gravitational *field* is always an approximation. In the case of weak gravity, like in everyday life, the approximation is very good. Many relativistic effects can be described with the gravitational field, without using space curvature or the metric tensor. In other words, instead of describing the complete space-time mattress, the gravitational field only describes the deviation of the mattress from the flat state, by pretending that the deviation is a separate entity, called the gravitational field. But what is the relativistically correct way to describe the gravitational field?

In a relativistic description of electrodynamics, the electromagnetic field has an elec- Page 497 tric and a magnetic component. The electric field is responsible for the inverse-square Coulomb force. In the same way, a relativistic description of (weak) gravity,* the gravitational field has an gravitoelectric and a gravitomagnetic component. The gravitoelectric field is responsible for the inverse square acceleration of gravity; what we call the gravitational field in everyday life is the gravitoelectric part of the full relativistic gravitational field.

In nature, all components of the mass-energy tensor produce gravity effects. In other words, not only mass and energy produce a field, but also mass or energy *currents*. This latter case is called gravitomagnetism (or frame dragging). The name is due to the analogy with electrodynamics, were not only charge density produces a field (the electric field), but also charge current (the magnetic field).

In the case of electromagnetism, the split between magnetic and electric field depends on the observer; each of the two can (partly) be transformed into the other. Exactly the Ref. 334 same happens in gravitation. Electromagnetism provides a good indication as to how the two types of gravitational fields behave; this intuition can directly be transferred to gravity. In electrodynamics, the acceleration of a charged particle is described by the Lorentz Page 484 equation

$$m\ddot{\mathbf{x}} = q\mathbf{E} - q\dot{\mathbf{x}} \times \mathbf{B} . \tag{248}$$

In other words, the change of *speed* is due to electric fields **E**, whereas magnetic fields **B** are those fields which give a velocity-dependent change of the *direction* of velocity, without changing the speed itself. Both changes depend on the value of the charge q. In the case of gravity this expression becomes

$$m\ddot{\mathbf{x}} = m\mathbf{G} - m\dot{\mathbf{x}} \times \mathbf{H} . \tag{249}$$

The role of charge is taken by mass. In this expression we already know the field **G**, given

* This section can be skipped at first reading.
* The approximation requires slow velocities, weak fields, as well as localized and stationary mass-energy distributions.

by

$$\mathbf{G} = \nabla\varphi = \nabla\frac{GM}{r} = -\frac{GM\mathbf{x}}{r^3} . \qquad (250)$$

As usual, the quantity φ is the (scalar) potential. The field $\mathbf{G}$ is the usual gravitational field of universal gravity, produced by every mass, and in this context is called the *gravitoelectric field*. Masses are the sources of the gravitoelectric field. The gravitoelectric field obeys $\Delta\mathbf{G} = -4\pi G\rho$. A *static* field $\mathbf{G}$ has no vortices; it obeys $\Delta \times \mathbf{G} = 0$.

It is not hard to show that if gravito*electric* fields exist, gravito*magnetic* fields must exist as well; the latter appear whenever one changes from an observer at rest to a moving one. (We will use the same argument in electrodynamics.) A particle falling perpendicularly towards an infinitely long rod already makes the point, as shown in Figure 178. An observer at rest with respect to the rod can describe the whole situation with gravitoelectric forces alone. A second observer, moving along the rod with constant speed, observes that the momentum of the particle *along the rod* also increases. This observer will thus not only measure a gravitoelectric field; he also measures a gravitomagnetic field. Indeed, a mass moving with velocity v produces a gravitomagnetic (3-) acceleration on a test mass m given by

Ref. 335
Page 497

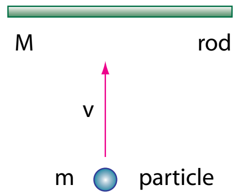


Figure 178 The reality of gravitomagnetism

Challenge 723 ny

$$m\mathbf{a} = -m\mathbf{v} \times \mathbf{H} \qquad (251)$$

where, *almost* as in electrodynamics, the static *gravitomagnetic field* $\mathbf{H}$ obeys

$$\mathbf{H} = \nabla \times \mathbf{A} = 16\pi N\rho\mathbf{v} \qquad (252)$$

where ρ is mass density of the source of the field and N is a proportionality constant. The quantity $\mathbf{A}$ is obviously called the *gravitomagnetic vector potential*. In nature, there are no sources for the gravitomagnetic field; the gravitomagnetic field thus obeys $\nabla\mathbf{H} = 0$.

Challenge 724 ny

When the situation in Figure 178 is evaluated, we find that the proportionality constant N is given by

$$N = \frac{G}{c^2} = 7.4 \cdot 10^{-28}\ \mathrm{m/kg} , \qquad (253)$$

an extremely small value. We thus find that like in the electrodynamic case, the gravitomagnetic field is weaker than the gravitoelectric field by a factor of c^2. It is thus hard to observe. In addition, a second aspect renders the observation of gravitomagnetism even more difficult. In contrast to electromagnetism, in the case of gravity there is no way to observe *pure* gravitomagnetic fields (why?); they are always mixed with the usual, gravitoelectric ones. For these reasons, gravitomagnetic effects have been measured for the first time only in the 1990s. We see that universal gravity is the approximation of general relativity appearing when all gravitomagnetic effects are neglected.

Challenge 725 n

In summary, *if a mass moves, it also produces a gravitomagnetic field*. How can one imagine gravitomagnetism? Let's have a look at its effects. The experiment of Figure 178 showed that a test mass is dragged along a moving rod. In our analogy of the vacuum as a mattress, it looks as if a moving rod drags the vacuum along with it, as well as any test mass that happens to be in that region. Gravitomagnetism can thus be seen as vacuum dragging. Due to a still widespread reticence of thinking vacuum as a mattress, the expression *frame dragging* is used instead.

In this description, *all frame dragging effects are gravitomagnetic effects*. In particular, a gravitomagnetic field also appears when a large mass *rotates*, as in the Lense–Thirring effect of Figure 176. For an angular momentum J the gravitomagnetic field $\mathbf{H}$ is a dipole field; it is given by

$$\mathbf{H} = \nabla \times \mathbf{h} = \nabla \times \left(-2 \frac{\mathbf{J} \times \mathbf{x}}{r^3} \right) \tag{254}$$

exactly as in the electrodynamic case. The gravitomagnetic field around a spinning mass has three effects.

First of all, like in electromagnetism, a spinning test particle with angular momentum $\mathbf{S}$ feels a *torque* if it is near a large spinning mass with angular momentum $\mathbf{J}$. And obviously, this torque $\mathbf{T}$ is given by

$$\mathbf{T} = \frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = \tfrac{1}{2}\, \mathbf{S} \times \mathbf{H} \,. \tag{255}$$

The torque leads to the precession of gyroscopes. For the Earth, this effect is extremely small: at the North Pole, the precession has a conus angle of 0.6 milli arc-seconds and a rotation rate of the order of 10^{-10} of the rotation of the Earth.

Since for a torque one has $\mathbf{T} = \dot{\mathbf{\Omega}} \times \mathbf{S}$, the dipole field a large rotating mass with angular momentum $\mathbf{J}$ yields a second effect. An orbiting mass will experience precession of its orbit plane. Seen from infinity one gets, for an orbit with semimajor axis a and eccentricity

Challenge 726 ny e,

$$\dot{\mathbf{\Omega}} = -\frac{\mathbf{H}}{2} = -\frac{G}{c^2}\frac{\mathbf{J}}{|\mathbf{x}|^3} + \frac{G}{c^2}\frac{3(\mathbf{J}\mathbf{x})\mathbf{x}}{|\mathbf{x}|^5} = \frac{G}{c^2}\frac{2\mathbf{J}}{a^3(1-e^2)^{3/2}} \tag{256}$$

which is the prediction by Lense and Thirring.* The effect is extremely small, giving a change of only 8 ″ per orbit for a satellite near the surface of the Earth. Despite this smallness and a number of larger effects disturbing it, Ciufolini's team managed to confirm the

Ref. 331 result.

As a third effect, a rotating mass leads to the precession of the periastron. This is a similar effect to the one that space curvature has on orbiting masses even if the central body does not rotate. The rotation in fact reduces the precession due to space-time curvature. This effect has been fully confirmed for the famous binary pulsar PSR B1913+16, as well as for the 'real' double pulsar PSR J0737-3039, discovered in 2003. This latter system shows a periastron precession of 16.9°/a, the largest known so far.

The use of the split into gravitoelectric and gravitomagnetic effects is an approximation to the description of gravity. Despite this limitation, the approach is useful. For example, it helps to answer questions such as: How can gravity keep the Earth around the

Challenge 728 ny Sun, if gravity needs 8 minutes to get from the Sun to us? To find the answer, thinking about the electromagnetic analogy can help. In addition, the split of the gravitational field

Challenge 727 ny * A homogeneous spinning sphere has an angular momentum given by $J = \frac{2}{5} M \omega R^2$.

into gravitoelectric and gravitomagnetic components also allows a simple description of gravitational waves.

Gravitational waves

One of the most fantastic predictions of physics is the existence of gravitational waves. Gravity waves* prove that empty space itself has the ability to move and vibrate. The basic idea is simple. Since space is elastic, like a large mattress in which we live, space should be able to oscillate in the form of propagating waves, like a mattress or any other elastic medium.

Table 34 The expected spectrum of gravitational waves

Frequency	Wavelength	Name	Expected appearance
$< 10^{-4}$ Hz	> 3 Tm	extremely low frequencies	slow binary star systems, supermassive black holes
10^{-4} Hz-10^{-1} Hz	3 Tm-3 Gm	very low frequencies	fast binary star systems, massive black holes, white dwarf vibrations
10^{-1} Hz-10^{2} Hz	3 Gm-3 Mm	low frequencies	binary pulsars, medium and light black holes
10^{2} Hz-10^{5} Hz	3 Mm-3 km	medium frequencies	supernovae, pulsar vibrations
10^{5} Hz-10^{8} Hz	3 km-3 m	high frequencies	unknown; possibly human made sources
$> 10^{8}$ Hz	< 3 m		unknown, possibly cosmological sources

Ref. 336 Jørgen Kalckar and Ole Ulfbeck have given a simple argument for the necessity of gravitational waves based on the existence of a maximum speed. They studied two equal masses falling towards each other due to gravitational attraction and imagined a spring between them. Such a spring will make the masses bounce towards each other again and again. The central spring stores the kinetic energy from the falling masses. The energy value can be measured by determining the length by which the spring is compressed. When the spring expands again and hurls the masses back into space, the gravitational attraction will gradually slow down the masses, until they again fall towards each other, thus starting the same cycle again.

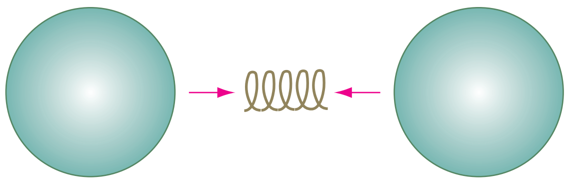


Figure 179 A Gedanken experiment showing the necessity of gravity waves

* To be strict, gravitational waves and gravity waves are different things; *gravity waves* are the surface waves of the sea, where gravity is the restoring force. However, in general relativity, both expressions are used interchangeably to mean the same thing.

However, the energy stored in the spring must get smaller with each cycle. Whenever a sphere detaches from the spring, it obviously is decelerated by the other sphere due to the gravitational attraction. Now comes the point. The value of this deceleration depends on the distance to the other mass; but since there is a maximal propagation velocity, the effective deceleration is given by the distance the other mass *had* when its gravity effect left towards the second mass. For two masses departing from each other, the effective distance is thus somewhat smaller than the momentary distance. In short, while departing, the real deceleration is *larger* than the one calculated without taking the time delay into account.

Similarly, when a mass falls back down towards the other, it is accelerated by the other mass according to the distance it had when the gravity effect started moving towards the other mass. Therefore, while approaching, the acceleration is *smaller* than the one calculated without time delay.

As a total effect, the masses arrive with a *smaller* energy than they departed with. At every bounce, the spring is compressed a little less. The difference of these two energies is lost by each mass; it is taken away by space-time. The energy difference is radiated away as gravitational radiation. The same story is told by mattresses. We remember that a mass deforms the space around it like a metal ball on a mattress deforms the surface around it. However, in contrast to actual mattresses, there is no friction between the ball and the mattress. If two metal balls continuously bang onto each other and then depart again, until they come back together, they will send out surface waves on the mattress. As we will see shortly, this effect has already been measured, with the difference that the two masses, instead of being reflected by a spring, were orbiting each other.

A simple mathematical description of gravity waves appears when the split into grav-
Ref. 337 itomagnetic and gravitoelectric effects is used. It does not take much effort to extend gravitomagnetostatics and gravitoelectrostatics to *gravitodynamics*. Just as electrodynamics can be deduced from Coulomb's attraction when one switches to other inertial observers,
Challenge 729 ny gravitodynamics can be deduced from universal gravity. One gets the four equations

$$\nabla \mathbf{G} = -4\pi G \rho \quad , \quad \nabla \times \mathbf{G} = -\frac{\partial \mathbf{H}}{\partial t}$$
$$\nabla \mathbf{H} = 0 \quad , \quad \nabla \times \mathbf{H} = -16\pi G \rho \mathbf{v} + \frac{N}{G}\frac{\partial \mathbf{G}}{\partial t} \ . \qquad (257)$$

We know two equations from above. The two other equations are expanded versions of what we encountered, taking time-dependence into account. Except for a factor 16 instead of 4 in the last equation, the equations for gravtitodynamics are the same as Maxwell's equations for electrodynamics.* These equations have a simple property: in vacuum, one

* The additional factor reflects the property that the ratio between angular momentum to energy (the 'spin') of gravity waves is different from that of electromagnetic waves. Gravity waves have spin 2, whereas electromagnetic waves have spin 1. We note that since gravity is universal, there can exist only a *single* kind of spin 2 radiation particle in nature. This is in strong contrast to the spin 1 case, of which there are several examples in nature.

By the way, the spin of radiation is a *classical* property. The *spin of a wave* is the ratio $E/L\omega$, where E is the energy, L the angular momentum, and ω is the angular frequency. For electromagnetic waves, this ratio is equal to 1, for gravitational waves, it is 2.

can deduce a *wave equation* for the gravitoelectric and the gravitomagnetic fields **G** and **H**. Challenge 730 ny (It is not hard; try!) In other words, *gravity can behave like a wave; gravity can radiate*. All this follows from the expression of universal gravity when applied to moving observers, with the requirement that neither observers nor energy can move faster than c. Both the above story with the spring and the present mathematical story use the same assumptions and arrive at the same conclusion.

Challenge 731 e A few manipulations show that the speed of these waves is given by

$$c = \sqrt{\frac{G}{N}} \; . \tag{258}$$

Page 520 This result corresponds to the electromagnetic expression

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \; . \tag{259}$$

The same letter has been used for the two speeds, as they are identical. Both influences travel with the speed common to all energy with vanishing rest mass. (We note that this is, strictly speaking, a prediction; the speed of gravitational waves has not yet been measured.) A claim from 2003 has turned out to be false. Ref. 338

How does one have to imagine these waves? We sloppily said above that a gravitational wave corresponds to a surface wave of the mattress; now we have to do better and imagine that we live *inside* the mattress. Gravitational waves are thus moving and oscillating deformations of the mattress, i.e., of space. Like in the case of mattress waves, it turns out that gravity waves are *transverse*. Thus they can be polarized. (Surface waves on mattresses cannot, because in two dimensions there is no polarization.) Gravity waves can be polarized in two independent ways. The effects of a gravitational wave are shown in

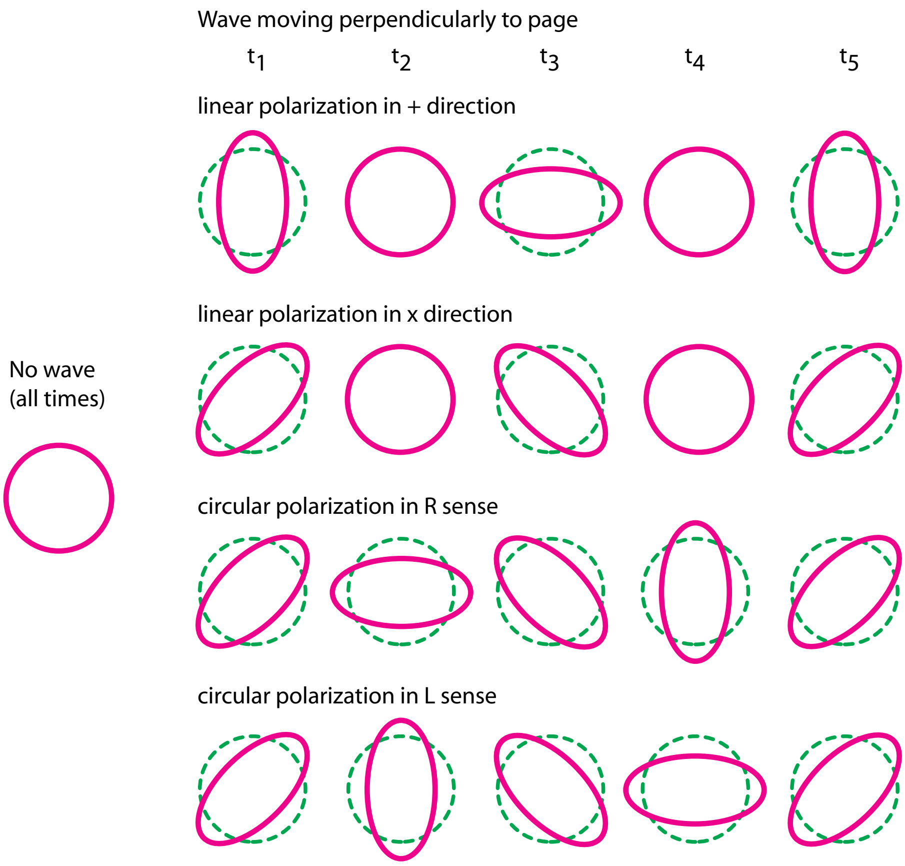


Figure 180 Effects on a circular or spherical body by a plane gravitational wave moving vertically to the page

Figure 180, both for linear and circular polarization.* We note that the waves are invariant under a rotation by π and that the two linear polarizations differ by an angle $\pi/4$; this shows that the particles corresponding to the waves, the gravitons, are of spin 2. (In

* A (small amplitude) plane gravity wave travelling in z-direction is described by a metric g given by

$$g = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1+h_{xx} & h_{xy} & 0 \\ 0 & h_{xy} & -1+h_{xx} & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{260}$$

where its two components, whose amplitude ratio determine the polarization, are given by

$$h_{ab} = B_{ab} \sin(kz - \omega t + \varphi_{ab}) \tag{261}$$

as in all plane harmonic waves. The amplitudes B_{ab}, the frequency ω and the phase φ are determined by the specific physical system. The general dispersion relation for the wave number k resulting from the wave equation is

$$\frac{\omega}{k} = c \tag{262}$$

and shows that the waves move with the speed of light.

general, the classical radiation field for a spin S particle is invariant under a rotation by $2\pi/S$. In addition, the two orthogonal linear polarizations of a spin S particle form an angle $\pi/2S$. For the photon, for example, the spin is 1; indeed, its invariant rotation angle is 2π and the angle formed by the two polarizations is $\pi/2$.)*

If we image empty space as a mattress that *fills* space, gravitational waves are wobbling deformations of the mattress. More precisely, Figure (180) shows that a wave of circular polarization has the same properties as a cork screw advancing through the mattress. We will discover later on why the analogy between a cork screw and a gravity wave with circular polarization works so well. Indeed, in the third part we will find a specific model of the space-time mattress material that automatically incorporates cork screw waves (instead of the spin 1 waves shown by everyday latex mattresses).

How does one produce gravitational waves? Obviously, masses must be accelerated. But how exactly? The conservation of energy does not allow that mass monopoles vary in strength. We also know form universal gravity that a spherical mass whose radius oscillates would not emit gravitational waves. In addition, the conservation of momentum does not allow changing mass dipoles. As a result, only *changing quadrupoles* can emit waves. For example, two masses in orbit around each other will emit gravitational waves. Also any rotating object which is not cylindrically symmetric around the rotation axis will do so. As a result, rotating an arm leads to gravitational wave emission. Most of these connections also apply for masses in mattresses. Are you able to point out the differences?

Challenge 732 ny

Challenge 733 ny

Einstein found that the amplitude h of waves at a distance r from a source is given to good approximation by the second derivative of the retarded quadrupole moment Q:

Ref. 339

$$h_{ab} = \frac{2G}{c^4}\frac{1}{r}d_{tt}Q^{\text{ret}}_{ab} = \frac{2G}{c^4}\frac{1}{r}d_{tt}Q_{ab}(t - r/c) . \qquad (264)$$

The expression shows that the amplitude of gravity waves *decreases only with* $1/r$, in contrast to naive expectations. However, this feature is the same as for electromagnetic waves. In addition, the small value of the prefactor, $1.6 \cdot 10^{-44}$ Wm/s, shows that truly gigantic systems are needed to produce quadrupole moment changes that yield any detectable length variations in bodies. To be convinced, just insert a few numbers, keeping in mind that the best presentdetectors are able to measure length changes down to $h = \delta l/l = 10^{-19}$. The production of detectable gravitational waves by humans is most probably impossible.

Challenge 734 ny

Gravitational waves, like all other waves, transport energy.* We specialize the general

In another gauge, a plane wave can be written as

$$g = \begin{pmatrix} c^2(1+2\varphi) & A_1 & A_2 & A_3 \\ A_1 & -1+2\varphi & h_{xy} & 0 \\ A_2 & h_{xy} & -1+h_{xx} & 0 \\ A_3 & 0 & 0 & -1 \end{pmatrix} \qquad (263)$$

where φ and $\mathbf{A}$ are the potentials such that $\mathbf{G} = \nabla\varphi - \frac{\partial \mathbf{A}}{c\partial t}$ and $\mathbf{H} = \nabla \times \mathbf{A}$.

* We note that since gravity is universal, there can exist only a *single* kind of spin-2 radiation particle in nature. This is in strong contrast to the spin-1 case, of which there are several examples in nature.

* Gravitomagnetism and gravitoelectricity, as in electrodynamics, allow one to define a gravitational Poynting vector. It is as easy to define and use as in the electrodynamic case.

Ref. 335

formula for the emitted power P to the case of two masses m_1 and m_2 in circular orbits
Ref. 302 around each other at distance l and get

$$P = -\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{G}{45c^5}\dddot{Q}^{\mathrm{ret}}_{ab}\dddot{Q}^{\mathrm{ret}}_{ab} = \frac{32}{5}\frac{G}{c^5}\left(\frac{m_1 m_2}{m_1+m_2}\right)^2 l^4 \omega^6 \qquad (265)$$

which, using Kepler's relation $4\pi^2 r^3/T^2 = G(m_1+m_2)$, becomes

$$P = \frac{32}{5}\frac{G^4}{c^5}\frac{(m_1 m_2)^2(m_1+m_2)}{l^5} . \qquad (266)$$

Ref. 302 For elliptical orbits, the rate increases with the ellipticity, as explained by Goenner. Inserting the values in the case of the Earth and the Sun, we get a power of about 200 W, and a value of 400 W for the Jupiter–Sun system. These values are so small that their effect cannot be detected at all.

For all orbiting systems, the frequency of the waves is twice the orbital frequency, as Challenge 735 ny you might want to check. These low frequencies make it even more difficult to detect them.

As a result, the only observation of effects of gravitational waves to date isin binary pulsars. Pulsars are small but extremely dense stars; even with a mass equal to that of the Sun, their diameter is only about 10 km. Therefore they can orbit each other at small distances and high speeds. Indeed, in the most famous binary pulsar system, PSR 1913+16, the two stars orbit each other in an amazing 7.8 h, even though their semimajor axis is about 700 Mm, just less than twice the Earth–Moon distance. Since their orbital speed is up to 400 km/s, the system is noticeably relativistic.

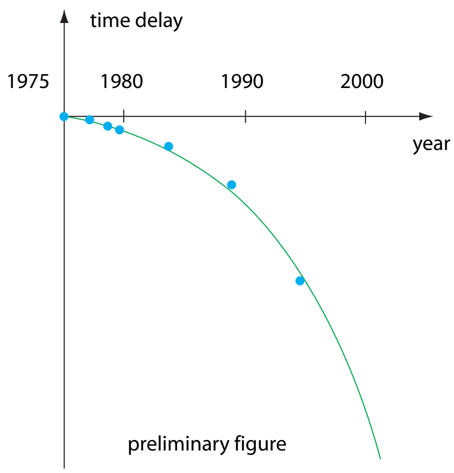


Figure 181 Comparison between measured time delay in the periastron of the binary pulsar psr 1913+16 and the prediction due to energy loss by gravitational radiation

Pulsars have a useful property: due to their rotation, they emit extremely regular radio pulses (hence their name), often in millisecond periods. Therefore it is easy to follow their orbit by measuring the change of pulse arrival time. In a famous experiment, a team of astrophysicists led by Joseph Taylor* measured the speed decrease of the binary pulsar system just Ref. 341 mentioned. Eliminating all other effects and collecting data for 20 years, they found a slowing down of the orbital frequency shown in Figure 181. The slowdown is due to gravity wave emission. The results exactly fit the prediction by general relativity, *without any* Challenge 736 ny *adjustable parameter*. (You might want to check that the effect must be quadratic in time.) Page 360 This is the only case so far that general relativity has been tested up to $(v/c)^5$ precision. To get an idea of the precision, this experiment detected a reduction of the orbit diameter

* In 1993 he shared the Nobel prize in physics for his life's work.

Figure to be included

Figure 182 Detection of gravitational waves

Ref. 341 of 3.1 mm per orbit, or 3.5 m per year! The measurements were possible only because the two stars in this system are neutron stars with small size, large velocities and purely gravitational interactions. The pulsar rotation period around its axis, about 59 ms, is known with eleven digits of precision, the orbital time of 7.8 h is known to ten digits and the Ref. 302 eccentricity of the orbit with 6 digits.

The *direct* detection of gravitational waves is one of the aims of experimental general relativity. The race is on since the 1990s. The basic idea is simple and taken from Figure 182: take four bodies for which the line connecting one pair is perpendicular to the line connecting the other pair. Then measure the distance changes of each pair. If a gravitational wave comes by, one pair will increase in distance and the other will decrease, at the *same* time.

Since gravitational waves cannot be produced in sufficient strength by humans, wave detection first of all requires the patience to wait for a strong enough wave to come by. Secondly, a system able to detect length changes of the order of 10^{-22} or better is needed – in other words, a lot of money. Any detection is guaranteed to make the news in television.*

It turns out that even for a body around a black hole, only about 6 % of the rest mass can be radiated away as gravitational waves; in particular, most of the energy is radiated during the final fall into the black hole, so that only quite violent processes, such as black hole collisions, are good candidates for detectable gravity wave sources.

Gravitational waves are a fascinating process. They still provide many topics to explore. Challenge 738 r For example: can you find a method to measure their speed? A well-publicized false claim Ref. 338 appeared in 2003. Indeed, a correct measurement that does not use the mentioned detectors would be a scientific sensation.

For the time being, another question on gravitational waves remains open: If all change is due to motion of particles, as the Greeks maintained, how do gravity waves fit into the Challenge 739 ny picture? If gravitational waves were made of particles, space-time would also have to be. We have to wait until the beginning of the third part of our ascent to say more.

Ref. 342 * The topic of gravity waves is full of interesting sidelines. For example, can gravity waves be used to power Challenge 737 ny a rocket? Yes, say Bonnor and Piper. You might ponder the possibility yourself.

Bending of light and radio waves

As we know from above, gravity also influences the motion of light. A far away observer measures a changing value for the light speed v near a mass. (Measured at the spot, the speed of light is of course always c.) It turns out that a far away observer measures a *lower* speed, so that for him, gravity has the same effects as a dense optical medium. It takes only a little bit of imagination that this effect will thus *increase* the bending of light near masses already deduced in 1801 by Soldner for universal gravity.

To calculate the effect, a simple way is the following. As usual, we use the coordinate system of flat space-time at infinity. The idea is to do all calculations to first order, as the value of the bending is very small.

Ref. 343 The angle of deflection α, to first order, is simply

$$\alpha = \int_{-\infty}^{\infty} \frac{\partial v}{\partial x} \mathrm{d}y \,, \tag{267}$$

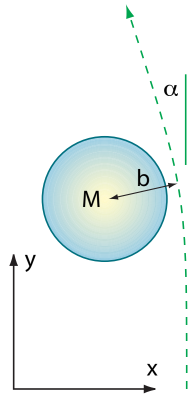


Figure 183 Calculating the bending of light by a mass

where v is the speed of light measured by a distant observer. (Can you Challenge 740 ny confirm it?) The next step is to use the Schwarzschild metric

$$\mathrm{d}\tau^2 = (1 - \frac{2GM}{rc^2})\mathrm{d}t^2 - \frac{\mathrm{d}r^2}{(c^2 - \frac{2GM}{r})} - \frac{r^2}{c^2}\mathrm{d}\varphi^2 \tag{268}$$

Challenge 741 ny and transform it into (x, y) coordinates to first order. That gives

$$\mathrm{d}\tau^2 = (1 - \frac{2GM}{rc^2})\mathrm{d}t^2 - (1 + \frac{2GM}{rc^2})\frac{1}{c^2}(\mathrm{d}x^2 + \mathrm{d}y^2) \tag{269}$$

which again to first order leads to

$$\frac{\partial v}{\partial x} = (1 - \frac{2GM}{rc^2})c \,. \tag{270}$$

It confirms what we know already, namely that far away observers see light *slowed down* when passing near a mass. Thus we can also speak of a height dependent index of refraction. In other words, constant *local* light speed leads to a *global* slowdown. This effect will play a role again shortly.

Challenge 742 ny Inserting the last result in (267) and using a smart substitution, we get a deviation angle α given by

$$\alpha = \frac{4GM}{c^2}\frac{1}{b} \tag{271}$$

where the distance b is the so-called *impact parameter* of the approaching light beam. Page 123 The resulting deviation angle α is *twice* the result we found for universal gravity. For a beam just above the surface of the Sun, the result is the famous value of $1.75''$ which was confirmed by the measurement expedition of 1919. (How did they measure the deviation Challenge 743 ny angle?) This was the experiment that made Einstein famous, as it showed that universal gravity is wrong. In fact, Einstein was lucky. Two earlier expeditions organized to measure the value had failed. In 1912, it was impossible to take data because of rain, and in 1914 in Crimea, scientists were arrested (by mistake) as spies, due to the beginning of the world war. But in 1911, Einstein had already published an *incorrect* calculation, giving only the Soldner value with half the correct size; only in 1915, when he completed general relativity, did he find the correct result. Therefore Einstein became famous only because of the failure of the two expeditions that took place before he published his correct calculation.

Ref. 344
Page 123

For high precision experiments around the Sun, it is more effective to measure the bending of radio waves, as they encounter fewer problems when they propagate through the solar corona. So far, over a dozen independent experiments did so, using radio sources in the sky which lie on the path of the Sun. They confirmed general relativity's prediction within a few per cent.

Ref. 324, Ref. 301, Ref. 302

So far, bending of radiation has also been observed near Jupiter, near certain stars, near several galaxies and near galaxy clusters. For the Earth, the angle is at most 3 nrad, too small to be measured yet, even though this may be feasible in the near future. There is a chance to detect this value if, as Andrew Gould proposes, the data of the satellite Hipparcos, which is taking precision pictures of the night sky, are analysed properly in the future.

Page 432

Of course, the bending of light also confirms that in a triangle, the sum of the angles does not add up to π, as is predicted later for curved space. (What is the sign of the curvature?)

Page 382
Challenge 744 ny

Time delay

The above calculation of the bending of light near masses shows that for a distant observer, light is slowed down near a mass. Constant *local* light speed leads to a *global* light speed slowdown. If light were not slowed down near a mass, it would go faster than c for an observer near the mass!* In 1964, Irwin Shapiro had the idea to measure this effect. He proposed two methods. The first was to send radar pulses to Venus, and measure the time for the reflection to get back to Earth. If the signals pass near the Sun, they will be delayed. The second was to use an artificial satellite communicating with Earth.

Ref. 345

The first measurement was published in 1968, and directly confirmed the prediction of general relativity within experimental errors. All subsequent tests have also confirmed the prediction within experimental errors, which nowadays are of the order of one part in a thousand. The delay has even been measured in binary pulsars, as there are a few such systems in the sky for which the line of sight lies almost precisely in the orbital plane.

Ref. 346
Ref. 347

The simple calculations presented here yield a challenge: Is it also possible to describe *full* general relativity – thus gravitation in *strong* fields – as a change of the speed of light with position and time induced by mass and energy?

Challenge 746 ny

* A nice exercise is to show that the bending of a slow particle gives the Soldner value, whereas with increasing speed, the value of the bending approaches twice that value. In all these considerations, the rotation of the mass has been neglected. As the effect of frame dragging shows, rotation also changes the deviation angle; however, in all cases studied so far, the influence is below the detection threshold.

Challenge 745 e

Effects on orbits

Astronomy allows the most precise measurements of motions, so that Einstein first of all tried to apply his results to the motion of planets. He thus looked for deviations of their motions from the predictions of universal gravity. Einstein found such a deviation: the precession of the perihelion of Mercury. The effect is shown in Figure 184. Einstein said later that the moment he found out that his calculation for the precession of Mercury matched observations was one of the happiest moments of his life.

The calculation is not difficult. In universal gravity, orbits are calculated by setting $a_{\text{grav}} = a_{\text{centri}}$, in other words, by setting $GM/r^2 = \omega^2 r$ and fixing energy and angular momentum. The mass of the orbiting satellite does not appear explicitly.

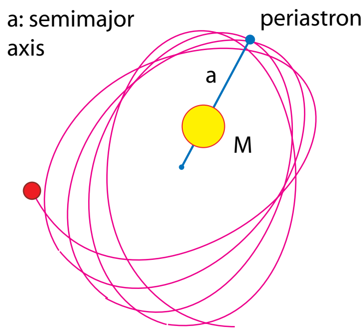


Figure 184 The orbit around a central body in general relativity

In general relativity, the mass of the orbiting satellite is made to disappear by rescaling energy and angular momentum as $e = E/mc^2$ and $j = J/m$. Next, the space curvature needs to be included. We use the Schwarzschild metric (268) mentioned above to deduce that the initial condition for the energy e, together with its conservation, leads to a relation between proper time τ and time t at infinity:

Ref. 301, Ref. 302
Page 351
Challenge 747 e

$$\frac{dt}{d\tau} = \frac{e}{1 - 2GM/rc^2} , \tag{272}$$

whereas the initial condition on the angular momentum j and its conservation implies that

$$\frac{d\varphi}{d\tau} = \frac{j}{r^2} . \tag{273}$$

These relations are valid for any particle, whatever its mass m. Inserting all this into the Schwarzschild metric, we get that the motion of a particle follows

$$\left(\frac{dr}{c\,d\tau}\right)^2 + V^2(j,r) = e^2 \tag{274}$$

where the effective potential V is given by

$$V^2(J,r) = \left(1 - \frac{2GM}{rc^2}\right)\left(1 + \frac{j^2}{r^2c^2}\right) . \tag{275}$$

Challenge 748 ny
Challenge 749 e

The expression differs slightly from the one in universal gravity, as you might want to check. We now need to solve for $r(\varphi)$. For *circular* orbits we get *two* possibilities

$$r_\pm = \frac{6GM/c^2}{1 \pm \sqrt{1 - 12\left(\frac{GM}{cj}\right)^2}} \tag{276}$$

where the minus sign gives a stable and the plus sign an unstable orbit. If $cj/GM < 2\sqrt{3}$, no stable orbit exists; the object will impact the surface or, for a black hole, be swallowed. There is a stable circular orbit *only* if the angular momentum j is larger than $2\sqrt{3}GM/c$. We thus find that in general relativity there is a smallest stable circular orbit, in contrast to universal gravity. The radius of this smallest stable circular orbit is $6GM/c^2 = 3R_S$.

What is the situation for *elliptical* orbits? Setting $u = 1/r$ in (274) and differentiating, the equation for $u(\varphi)$ becomes

$$u' + u = \frac{GM}{j^2} + \frac{3GM}{c^2}u^2 . \tag{277}$$

Without the nonlinear correction due to general relativity on the far right, the solutions
Challenge 750 e
are the famous *conic sections*

$$u_0(\varphi) = \frac{GM}{j^2}(1 + \varepsilon \cos\varphi) \tag{278}$$

i.e. ellipses, parabolas or hyperbolas. The type of conic section depends on the value of the parameter ε, the so-called *eccentricity*. We know the shapes of these curves from universal
Page 119
gravity. Now, general relativity introduces the nonlinear term on the right hand side of equation (277). Thus the solutions are not conical sections any more; however, as the
Challenge 751 e
correction is small, a good approximation is given by

$$u_1(\varphi) = \frac{GM}{j^2}\left[1 + \varepsilon \cos\left(\varphi - \frac{3G^2M^2}{j^2c^2}\varphi\right)\right] . \tag{279}$$

The hyperbolas and parabolas of universal gravity are thus slightly deformed. Instead of elliptical orbits we get the famous rosetta path shown in Figure 184. Such a path is above all characterized by a periastron shift. The *periastron*, or *perihelion* in the case of the Sun, is the nearest point to the central body reached by an orbiting body. The periastron turns
Challenge 752 e
around the central body with an angle

$$\alpha \approx 6\pi \frac{GM}{a(1-\varepsilon^2)c^2} \tag{280}$$

for every orbit, where a is the *semimajor axis*. For Mercury, the value is $43''$ per century. Around 1900, this was the only known effect that was unexplained by universal gravity; when Einstein's calculation led him to exactly that value, he was overflowing with joy for many days.

To be sure about the equality between calculation and experiment, all other effects leading to rosetta paths must be eliminated. For some time, it was thought that the quadrupole moment of the Sun could be an alternative source of this effect; later measurements ruled out this possibility.

In the meantime, the perihelion shift has been measured also for the orbits of Icarus, Venus and Mars around the Sun, as well as for several binary star systems. In binary
Ref. 347
pulsars, the periastron shift can be as large as several degrees per year. In all cases, expres-

sion (280) describes the motion within experimental errors.

We note that even the rosetta orbit itself is not really stable, due to the emission of gravitational waves. But in the solar system, the power lost this way is completely negligible even over thousands of millions of years, as we saw above, so that the rosetta path remains a good description of observations. Page 372

The geodesic effect

When a pointed body orbits a central mass m at distance r, the direction of the tip will not be the same after a full orbit. This effect exists only in general relativity. The angle α describing the direction change is given by

$$\alpha = 2\pi\left(1-\sqrt{1-\frac{3Gm}{rc^2}}\right) \approx \frac{3\pi Gm}{rc^2} \; . \qquad (281)$$

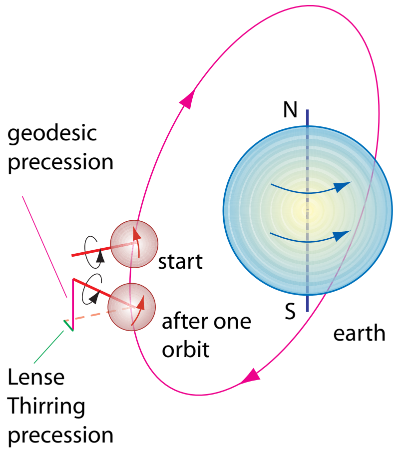


Figure 185 The geodesic effect

The angle change is called the *geodesic effect* – 'geodetic' in other languages. It is a further consequence of the split into gravitoelectric and gravitomagnetic fields, as you may want to show. Obviously, it does not exist in universal gravity. Challenge 753 e

In the case that the pointing of the orbiting body is realized by an intrinsic rotation, such as for a spinning satellite, the geodesic effect produces a *precession* of the axis. Thus the effect is comparable to spin-orbit coupling in atomic theory. (The Lense–Thirring effect mentioned above is analogous to spin-spin coupling.)

Ref. 348 The geodesic effect, or geodesic precession, was predicted by Willem de Sitter in 1916; in particular, he proposed to detect that the Earth–Moon system would change its pointing direction in its fall around the Sun. The effect is tiny; for the axis of the Moon the precession angle is about 0.019 arcsec per year. The effect was first detected in 1987 by an Ref. 349 Italian team for the Earth–Moon system, through a combination of radio-interferometry and lunar ranging, making use of the Cat's-eyes deposited by the cosmonauts on the Moon. Experiments to detect it in artificial satellites are also under way.

At first sight, geodesic precession is similar to the Thomas precession found in special Page 282 relativity. In both cases, a transport along a closed line results in the loss of the original direction. However, a careful investigation shows that Thomas precession can be *added* to geodesic precession by applying some additional, non-gravitational interaction, so that the analogy is shaky.

We now terminate the discussion of weak gravity effects. We turn to strong gravity, where curvature cannot be neglected and where the fun is even more intense.

Curiosities about weak fields

Here are some issues to think about.

- Is there a static, oscillating gravitational field?

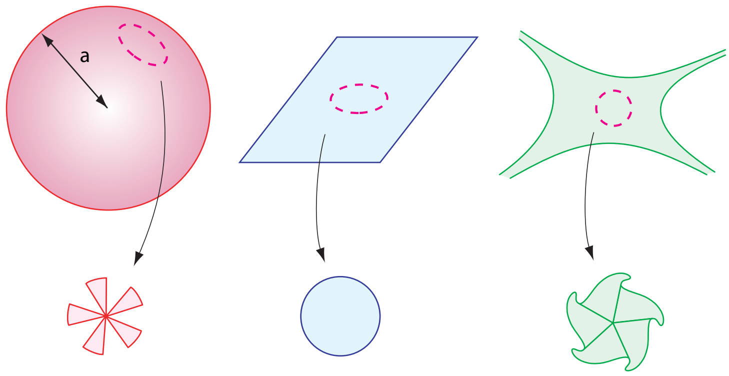


Figure 186 Positive, vanishing and negative curvature in two dimensions

- Are beams of gravitational waves, analogous to beams of light, possible?
- Would two parallel beams of gravitational waves attract each other?

How is curvature measured?

We saw that in the precise description of gravity, motion depends on space-time curvature. In order to add numbers to this idea, we first of all need to describe curvature itself as accurately as possible. To clarify the issue, we will start the discussion in two dimensions, and then go over to three and four dimensions.

Obviously, a flat sheet of paper has no curvature. If we roll it into a cone or a cylinder, it gets what is called *extrinsic curvature*; however, the sheet of paper still looks flat for any two-dimensional animal living on it – as approximated by an ant walking over it. In other words, the *intrinsic curvature* of the sheet of paper is zero even if the sheet as a whole is extrinsically curved. (Can a one-dimensional space have intrinsic curvature? Is a torus internally curved?) Challenge 754 n

Intrinsic curvature is thus the stronger concept, measuring the curvature which can be observed even by an ant. The surface of the Earth, the surface of an island, or the slopes of a mountain* are intrinsically curved. Whenever we talk about curvature in general relativity, we always mean *intrinsic* curvature, since any observer in nature is by definition in the same situation as an ant on a surface: their experience, their actions and plans always only concern their closest neighbourhood in space and time.

But how precisely can an ant determine whether it lives on an intrinsically curved surface?** One way is shown in Figure 186. The ant can check whether either the circumference of a circle or its area fits with the measured radius. She can even use the difference between the two numbers as a measure for the local intrinsic curvature, if she takes the limit for vanishingly small circles and if she normalizes the values correctly. In other words, the ant can imagine to cut out a little disk around the point she is on, to iron it flat and to check whether the disk would tear or produce folds. Any two-dimensional

Challenge 755 e

* Except if the mountain has the shape of a perfect cone. Can you confirm this?

** Note that the answer to this question also tells how to distinguish real curvature from curved coordinate systems on a flat space. This question is often put by those approaching general relativity for the first time.

surface is intrinsically curved whenever ironing is not sufficient to make a flat street map out of it.

This means that we can recognize intrinsic curvature also by checking whether two parallel lines stay parallel, approach each other, or depart from each other. In the first case, such as lines on a paper cylinder, the surface is said to have *vanishing* intrinsic curvature; a surface with approaching parallels, such as the Earth, is said to have *positive* curvature, and a surface with diverging parallels, such as a saddle, is said to have *negative* curvature. In short, positive curvature means that we are more restricted in our movements, negative that we are not. A *constant* curvature even implies to be locked in. You might want to check this with Figure 186.

The third way to measure curvature uses triangles. On curved surfaces the sum of angles in a triangle is either larger or smaller than π.

Ref. 350 Let us see how we can quantify curvature. First a question of vocabulary: a sphere with radius a is said, by definition, to have an intrinsic curvature $K = 1/a^2$. Therefore a plane has null curvature. You might check that for a circle on a sphere, the measured radius r, circumference C, and area A are related by (Challenge 756 e)

$$C = 2\pi r \left(1 - \frac{K}{6} r^2 + ...\right) \quad \text{and} \quad A = \pi r^2 \left(1 - \frac{K}{12} r^2 + ...\right) \tag{282}$$

where the dots imply higher order terms. This allows one to define the intrinsic curvature K, also called the *Gaussian* curvature, for a general point on a two-dimensional surface in either of the following two equivalent ways:

$$K = 6 \lim_{r \to 0} \left(1 - \frac{C}{2\pi r}\right) \frac{1}{r^2} \quad \text{or} \quad K = 12 \lim_{r \to 0} \left(1 - \frac{A}{\pi r^2}\right) \frac{1}{r^2} \ . \tag{283}$$

This expression allows an ant to measure the intrinsic curvature at each point for any smooth surface.* From now on in this text, *curvature* will always mean *intrinsic* curvature. Note that the curvature can be different from place to place, and that it can be positive, like for an egg, or negative, like the inside of any torus. Also a saddle is an example for the latter case, but, unlike the torus, with a curvature *changing* from point to point. In fact, it is not possible at all to fit a surface of *constant* negative curvature inside three-dimensional space; one needs at least four dimensions, as you can find out if you try to imagine the situation. (Challenge 758 e)

For any surface, at *every* point, the direction of maximum curvature and the direction of minimum curvature are *perpendicular* to each other. This connection, shown in Figure 187, was discovered by Leonhard Euler in the eighteenth century. You might want to check this with a tea cup, with a sculpture by Henry Moore, or with any other curved object from your surroundings, such as a Volkswagen Beetle. The Gaussian curvature K (Challenge 759 e)

* If the n-dimensional volume of a sphere is written as $V_n = C_n r^n$ and the n-dimensional surface as $O_n = n C_n r^{n-1}$, one can generalize the expressions to (Ref. 351)

$$K = 3(n+2) \lim_{r \to 0} \left(1 - \frac{V_n}{C_n r^n}\right) \frac{1}{r^2} \quad \text{or} \quad K = 3n \lim_{r \to 0} \left(1 - \frac{O_n}{n C_n r^{n-1}}\right) \frac{1}{r^2} \ , \tag{284}$$

as shown by Vermeil. A famous riddle is to determine C_n. (Challenge 757 ny)

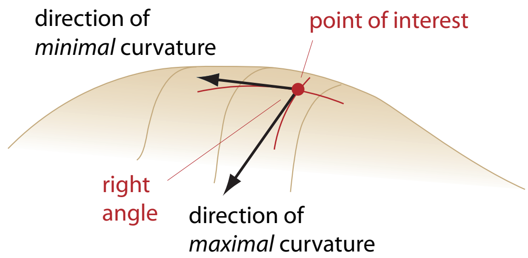


Figure 187 The maximum and minimum curvature of a curved surface

defined in (283) is in fact the product of the two corresponding inverse curvature radii. Thus, even though line curvature is not an intrinsic property, this special product is. Gaussian curvature is a measure for the instrinsic curvature. Intrinsic measures of curvature are needed if one is forced to stay inside the surface or space one is exploring. Physicists are thus particularly interested in Gaussian curvature and its higher-dimensional analogies.

For *three*-dimensional 'surfaces', the issue is a bit more involved. First of all, we have difficulties imagining the situation. But we can still visualize that the curvature of a small disk around a point will depend on its orientation. Let us first look at the simplest case. If the curvature at a point is the same in all directions, the point is called *isotropic*. We can imagine a small sphere around that point. In this special case, in three dimensions, the relation between the measured radius and the measured sphere surface A leads us to define the curvature K as

Challenge 760 ny

$$K = 9 \lim_{r \to 0} \Big(1 - \frac{A}{4\pi r^2}\Big) \frac{1}{r^2} = 18 \lim_{r \to 0} \frac{r - \sqrt{A/4\pi}}{r^3} = 18 \lim_{r \to 0} \frac{r_{\text{excess}}}{r^3} . \qquad (285)$$

Defining the *excess radius* as $r_{\text{excess}} = r - \sqrt{A/4\pi}$, we get that for a three-dimensional space, *the curvature is eighteen times the excess radius of a small sphere divided by the cube of its radius.* A positive curvature is equivalent to a positive excess radius, and similarly for vanishing and negative cases.

Of course, this value is only an average. The precise way requires to define curvature with disks; these values will *differ* from the values calculated by using the sphere, as they will depend on the *orientation* of the disk. However, all possible disk curvatures at a given point are related among each other and must form a tensor. (Why?) For a full description of curvature, we thus have to specify, as for any tensor in three dimensions, the main curvatures in three orthogonal directions.*

Challenge 761 ny

What are the curvature values in the space around us? Already in 1827, the mathematician and physicist Friedrich Gauß checked whether the three angles formed by three mountain peaks near his place of residence added up to π. Nowadays we know that the

* These three disk values are not independent however, since together, they must yield the just mentioned average volume curvature K. In total, there are thus *three* independent scalars describing the curvature in three dimensions (at each point). With the metric tensor g_{ab} and the Ricci tensor R_{ab} to be introduced below, one choice is to take for the three independent numbers the values $R = -2K$, $R_{ab}R^{ab}$ and $\det R/\det g$.

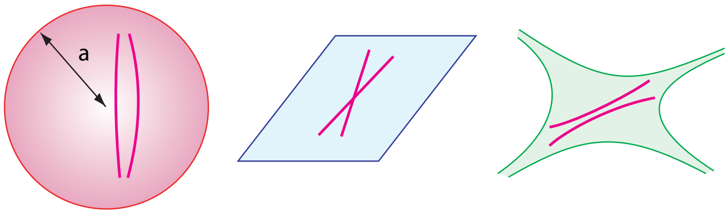


Figure 188 Curvature (in two dimensions) and geodesic behaviour

deviation δ from the angle π on the surface of a body of mass M and radius r is given by

$$\delta = \pi - (\alpha + \beta + \gamma) \approx A_{\text{triangle}} \frac{GM}{r^3 c^2} . \qquad (286)$$

This expression is typical for hyperbolic geometries. For the case of mathematical negative curvature, it was already deduced by Johann Lambert (1728–1777). However, only Einstein discovered that the negative curvature is related to the mass and gravitation of a body. For the case of the Earth and typical mountain distances, the angle δ is of the order of 10^{-14} rad. Gauss had no chance to detect any deviation, and in fact he detected none. Even today, studieswith lasers and high precision set-ups, no deviation has been detected yet – on Earth. The right-hand factor, which measures the curvature of space-time on the surface of the Earth, is simply too small. But Gauss did not know, as we do today, that gravity and curvature go hand in hand.

Curvature and space-time

> Notre tête est ronde pour permettre à la pensée de changer de direction.*
>
> Francis Picabia

In nature, with *four* space-time dimensions, specifying curvature requires a more involved approach. First of all, the use of space-time coordinates automatically introduces the speed of light c as limit speed, which is a central requirement in general relativity. Furthermore, the number of dimensions being four, we expect a value for an average curvature at a point, defined by comparing the 4-volume of a 4-sphere in space-time and with the one deduced from the measured radius; then we expect a set of 'almost average' curvatures defined by 3-volumes of 3-spheres in various orientations, plus a set of 'low-level' curvatures defined by usual 2-areas of usual 2-disks in even more orientations. Obviously, we need to bring some order in this set, and we need to avoid the double counting we already encountered in the case of three dimensions.

Fortunately, physics can help to make the mathematics easier. We start by defining what we mean by curvature of space-time. Then we will define curvatures for disks of various orientations. To achieve this, we translate the definition of curvature into another picture, which allows us to generalize it to time as well. Figure 188 shows that the curvature

* 'Our head is round in order to allow our thougths to change direction.' Francis Picabia (b. 1879 Paris, d. 1953 Paris) French dadaist and surrealist painter.

K also describes how geodesics *diverge*. Geodesics are the straightest paths on a surface, i.e. those paths that a tiny car or tricycle would follow if it drives on the surface keeping the steering wheel straight.

Challenge 762 e

If a space is curved, the separation s will increase along the geodesics as

$$\frac{d^2 s}{dl^2} = -Ks + \text{higher orders} \tag{287}$$

where l measures the length along the geodesic, and K is the curvature, in other words, the inverse squared curvature radius. In space-time, this relation is extended by substituting proper length with proper time (times the speed of light). Thus separation and curvature are related by

$$\frac{d^2 s}{d\tau^2} = -Kc^2 s + \text{higher orders} \ . \tag{288}$$

But this is the definition of an acceleration. In other words, what in the purely spatial case is described by *curvature*, in the case of space-time becomes the *relative acceleration* of two particles freely falling from nearby points. Indeed, we encountered these accelerations already: they describe tidal effects. In short, space-time curvature and tidal effects are precisely the same.

Page 120

Obviously, the value of tidal effects and thus of curvature will depend on the orientation – more precisely on the orientation of the space-time plane formed by the two particle velocities. The definition also shows that K is a tensor, so that later on we will have to add indices to it. (How many?) The fun is that we can avoid indices for a while by looking at a special combination of spatial curvatures. If we take three planes in space, all orthogonal to each other and through the same point, the *sum* of the three so-called *sectional* curvature values does *not* depend on the observer. (This corresponds to the tensor trace.) Can you confirm this, by using the definition of the curvature just given?

Challenge 763 ny

Ref. 352

Challenge 764 ny

The sum of three sectional curvatures defined for mutually orthogonal planes $K_{(12)}$, $K_{(23)}$ and $K_{(31)}$, is related to the excess radius defined above. Can you find out how?

Challenge 765 ny

If a surface has *constant* (intrinsic) curvature, i.e. the same curvature at all locations, geometrical objects can be moved around without deforming them. Can you picture this?

Challenge 766 e

In summary, curvature is not such a difficult concept. It describes the *deformation* of space-time. If we imagine space (-time) as a big blob of rubber in which we live, the curvature at a point describes how this blob is squeezed at that point. Since we live *inside* the rubber, we need to use 'insider' methods, such as excess radii and sectional curvatures, to describe the deformation. Relativity is only difficult to learn because people often do not like to think about the vacuum in this way, and even less to explain it in this way. (For a hundred years it was a question of faith for every physicist to say that the vacuum is empty.) Picturing vacuum as a substance can help imagination in many ways in understanding general relativity.

Curvature and motion in general relativity

As mentioned above, one half of general relativity is the statement that any object moves along paths of *maximum* proper time, i.e. along geodesics. The *other* half is contained in a single expression: the sum of all three *proper* sectional *spatial* curvatures at a point is given by

$$K_{(12)} + K_{(23)} + K_{(31)} = \frac{8\pi G}{c^4} W^{(0)} \tag{289}$$

where $W^{(0)}$ is the *proper* energy density at the point, *and* this statement is valid for *every* observer. The lower indices indicate the mixed curvatures defined by the three orthogonal directions 1, 2 and 3. This is all of general relativity in one paragraph.

Challenge 767 e

An equivalent way to describe the expression is easily found using the excess radius defined above, by introducing the mass $M = V W^{(0)}/c^2$. We get

$$r_{\text{excess}} = r - \sqrt{A/4\pi} = \frac{G}{3c^2} M \ . \tag{290}$$

In short, general relativity affirms that for every observer, *the excess radius of a small sphere is given by the mass inside the sphere.**

Note that the expression means that the average space curvature at a point in empty space *vanishes*. As we will see shortly, this means that near a spherical mass the curvature *towards* the mass and twice the curvature *around* the mass exactly compensate each other.

Curvature will also differ from point to point. In particular, the expression implies that if energy *moves*, curvature will move with it. In short, both space curvature and, as we will see shortly, space-time curvature *change* over space and time.

We note in passing that curvature has an annoying effect: the relative velocity of *distant* observers is undefined. Can you provide the argument? In curved space, relative velocity is defined only for *nearby* objects – in fact only for objects with no distance at all. Only in flat space are relative velocities of distant objects well defined.

Challenge 768 ny

The quantities appearing in expression (289) are *independent* of the observer. But often people want to use observer-dependent quantities. The relation then gets more involved; the single equation (289) must be expanded to ten equations, called *Einstein's field equations*. They will be introduced below. But before we do that, we check that general relativity makes sense. We skip the check that it contains special relativity as limiting case, and directly go to the main test.

Universal gravity

> The only reason which keeps me here is gravity.
> Anonymous

For small velocities, the temporal curvatures $K_{(0j)}$ of expression (287) turn out to have a special property. In this case, they can be defined as the second spatial derivatives of a single scalar function φ. In other words,

Challenge 769 e

Ref. 353

* Another, equivalent way is to say that for small radii the area A is given by

$$A = 4\pi r^2 (1 + \frac{1}{9} r^2 R) \tag{291}$$

where R is the Ricci scalar to be introduced later on.

$$K_{(0j)} = \frac{\partial^2 \varphi}{\partial (x^j)^2} \, . \tag{292}$$

In everyday situations, the function φ turns out to be the gravitational potential. Indeed, universal gravity is the description of general relativity for small speeds and small spatial curvature. These two limits imply, making use of $W^{(0)} = \rho c^2$ and $c \to \infty$, that

$$K_{(ij)} = 0 \quad \text{and} \quad \mathbf{K}_{(01)} + \mathbf{K}_{(02)} + \mathbf{K}_{(03)} = 4\pi G \rho \, . \tag{293}$$

In other words, for slow speeds, space is flat and the potential obeys Poisson's equation. Universal gravity is thus indeed the limit of general relativity.

Can you show that relation (289) between curvature and energy density indeed means that time near a mass depends on the height, as stated in the beginning of this chapter? Challenge 770 ny

The Schwarzschild metric

Ref. 352 What is the curvature of space-time near a spherical mass?

– CS – more to be inserted – CS –

Challenge 771 ny The curvature of the Schwarzschild metric is given by

$$\begin{aligned} K_{r\varphi} = K_{r\theta} &= -\frac{G}{c^2}\frac{M}{r^3} \quad \text{and} \quad K_{\theta\varphi} = 2\frac{G}{c^2}\frac{M}{r^3} \\ K_{t\varphi} = K_{t\theta} &= \frac{G}{c^2}\frac{M}{r^3} \quad \text{and} \quad K_{tr} = -2\frac{G}{c^2}\frac{M}{r^3} \end{aligned} \tag{294}$$

Ref. 352 Page 120 everywhere. The dependence on $1/r^3$ follows from the general dependence of all tidal effects; we have already calculated them in the chapter on universal gravity. The factors G/c^2 are due to the maximum force of gravity; only the numerical prefactors need to be calculated from general relativity. The average curvature obviously vanishes, as it does for all vacuum. As expected, the values of the curvatures near the surface of the Earth are exceedingly small. Challenge 772 ny

Curiosities and fun challenges about curvature

Every physicist should have an intuitive understanding about curvature.

▪ A fly has landed on the outside of a cylindrical glass, 1 cm below its rim. A drop of honey is located halfway around the glass, also on the outside, 2 cm below the rim. What is the shortest distance to the drop? What is the shortest distance if the drop is on the *inside* of the glass? Challenge 773 e

▪ Where are the points of highest and lowest Gaussian curvature on an egg? Challenge 774 e

– CS – more to come – CS –

All observers: heavier mathematics*

Jeder Straßenjunge in unserem mathematischen Göttingen versteht mehr von vierdimensionaler Geometrie als Einstein. Aber trotzdem hat Einstein die Sache gemacht, und nicht die großen Mathematiker.

David Hilbert**

Now that we have a feeling for curvature, we want to describe it in a way that allows *any* observer to talk to any *other* observer. Unfortunately, this means to use formulae with tensors. These formulae look exactly the way that non-scientists imagine: daunting. The challenge is to be able to see in each of the expressions the essential point (e.g. by forgetting all indices for a while) and not to be impressed by those small letters sprinkled all over them.

The curvature of space-time

Il faut suivre sa pente, surtout si elle monte.*

André Gide

We mentioned above that a 4-dimensional space-time is described by 2-curvature, 3-curvature and 4-curvature. Many texts on general relativity start with 3-curvature. These curvatures describing the distinction between the 3-volume calculated from a radius and the actual 3-volume. They are described by the *Ricci tensor*.** With an argument we encountered already for the case of geodesic deviation, it turns out that the Ricci tensor describes how the shape of a spherical cloud of freely falling particles is deformed on its path.

– CS – to be expanded – CS –

In short, the Ricci tensor is the general relativistic version of $\Delta\varphi$, or better, of $\Box\varphi$.

Obviously, the most global, but least detailed description of curvature is the one describing the distinction between the 4-volume calculated from a measured radius and the actual 4-volume. This is the *average curvature* at a space-time point and is described by the so-called *Ricci scalar R* defined as

$$R = -2K = -\frac{2}{r^2_{\text{curvature}}} \ . \tag{295}$$

It turns out that the Ricci scalar can be derived from the Ricci tensor by a so-called *contraction*, the name for the precise averaging procedure that is needed. For tensors of rank

* This section might be skipped at first reading. The section on cosmology, on page 402, is then the right point to continue.

** Every street urchin in our mathematical Göttingen knows more about four-dimensional geometry than Einstein. Nevertheless, it was Einstein who did the work, not the great mathematicians.

* 'One has to follow one's inclination, especially if it climbs upwards.

** It is named after the Italian mathematician Gregorio Ricci.

two, contraction is the same as the taking of the trace:

$$R = R^{\lambda}{}_{\lambda} = g^{\lambda\mu} R_{\lambda\mu} \ . \tag{296}$$

The Ricci scalar describes the curvature averaged over space *and* time. In the image of a falling spherical cloud, the Ricci scalar describes the volume change of the cloud. The Ricci scalar always vanishes in vacuum. This result allows one, on the surface of the Earth, to relate the spatial curvature values and the change of time with height.

Challenge 775 ny

Now comes one of the issues discovered by Einstein in two years of hard work. The quantity of importance for the description of curvature in nature is not the Ricci tensor R_{ab}, but a tensor built from it. This *Einstein tensor* G_{ab} is defined mathematically (for vanishing cosmological constant) as

$$G_{ab} = R_{ab} - \frac{1}{2} g_{ab} R \ . \tag{297}$$

It is not difficult to get its meaning. The value G_{00} is the sum of sectional curvatures in the planes *orthogonal* to the 0 direction and thus the sum of all spatial sectional curvatures:

$$G_{00} = K_{(12)} + K_{(23)} + K_{(31)} \ . \tag{298}$$

Similarly, the diagonal elements G_{ii} are the sum (taking into consideration the minus signs of the metric) of sectional curvatures in the planes *orthogonal* to the i direction. For example, we have

$$G_{11} = K_{(02)} + K_{(03)} - K_{(23)} \ . \tag{299}$$

The other components are defined accordingly. The distinction between the Ricci tensor and the Einstein tensor is thus the way in which the sectional curvatures are combined: disks *containing* the coordinate in question in one case, disks *orthogonal* to the coordinate in the other case. Both describe the curvature of space-time equally, and fixing one means fixing the other. (What is the trace of the Einstein tensor?)

Challenge 776 ny

The Einstein tensor is symmetric, which means that it has *ten* independent components. Most importantly, its divergence vanishes; it therefore describes a conserved quantity. And this was the key property which allowed Einstein to relate it to mass and energy in mathematical language.

The description of momentum, mass and energy

Obviously, for a complete description of gravity, also the motion of momentum and energy needs to be quantified in such a way that any observer can talk to any other. We have seen that momentum and energy always appear together in relativistic descriptions; the next step is thus to find out how this needs to be done in detail, for general observers.

First of all, the quantity describing energy, let us call it T, must be defined using the energy–momentum vector $\mathbf{p} = m\mathbf{u} = (\gamma mc, \gamma m\mathbf{v})$ of special relativity. Furthermore, T does not describe a single particle, but the way energy–momentum is distributed over space and time. As a consequence, it is most practical to use T to describe a *density* of energy and momentum. T will thus be a *field*, and depend on time and space, a fact usually written as $T = T(t, x)$.

Since the energy–momentum density T describes a density over space and time, it defines, at every space-time point and for every infinitesimal surface $d\mathbf{A}$ around that point, the flow of energy–momentum $d\mathbf{p}$ through that surface. In other words, T is defined by the relation

$$d\mathbf{p} = T\ d\mathbf{A}\ . \tag{300}$$

The surface is assumed to be characterized by its normal vector $d\mathbf{A}$. Since the energy–momentum density is a proportionality factor between two vectors, T is a *tensor*. Of course, we are talking about 4-flows and 4-surfaces here. Therefore the energy–momentum density tensor can be split in the following way:

$$T = \left(\begin{array}{c|ccc} w & S_1 & S_2 & S_3 \\ \hline S_1 & t_{11} & t_{12} & t_{13} \\ S_2 & t_{21} & t_{22} & t_{23} \\ S_3 & t_{31} & t_{32} & t_{33} \end{array}\right) = \left(\begin{array}{c|c} \text{energy} & \text{energy flow density, or} \\ \text{density} & \text{momentum density} \\ \hline \text{energy flow or} & \text{momentum} \\ \text{momentum density} & \text{flow density} \end{array}\right) \tag{301}$$

where $w = T_{00}$ is a 3-scalar, $\mathbf{S}$ a 3-vector and t a 3-tensor. The total quantity T is called the *energy–momentum (density) tensor*. It has two essential properties: it is symmetric and its divergence vanishes.

The vanishing divergence of the tensor T, often written as

$$\partial_a T^{ab} = 0 \quad \text{or abbreviated} \quad T^{ab}{}_{,a} = 0 \tag{302}$$

expresses that the tensor describes a *conserved* quantity. In every volume, energy can change only via flow through its boundary surface. Can you confirm that the description of energy–momentum with this tensor follows the requirement that any two observers, differing by position, orientation, speed *and* acceleration, can communicate their results to each other?

Challenge 777 ny

The energy–momentum density tensor gives a full description of the distribution of energy, momentum and mass over space and time. As an example, let us determine the energy–momentum density for a moving liquid. For a liquid of density ρ, a pressure p and a 4-velocity $\mathbf{u}$, we have

$$T^{ab} = (\rho_0 + p)u^a u^b - p g^{ab} \tag{303}$$

where ρ_0 is the density measured in the comoving frame, the so-called *proper* density.* Obviously, ρ, ρ_0 and p depend on space and time.

* In the *comoving* frame we thus have

$$T^{ab} = \begin{pmatrix} \rho_0 c^2 & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}. \tag{304}$$

Of course, for a particular material fluid, we need to know how pressure p and density ρ are related. A full material characterization thus requires the knowledge of the relation

$$p = p(\rho) . \tag{305}$$

This relation is a material property and thus *cannot* be determined from relativity. It has to be derived from the constituents of matter or radiation and their interactions. The simplest possible case is *dust*, i.e. matter made of point particles* with no interactions at all. Its energy–momentum tensor is given by

$$T^{ab} = \rho_0 u^a u^b . \tag{306}$$

Challenge 778 ny Can you explain the difference from the liquid case?

Challenge 779 ny The divergence of the energy–momentum tensor vanishes for all times and positions, as you may want to check. This property is the same as for the Einstein tensor presented above. But before we elaborate on the issue, a short remark. We did not take into account *gravitational energy*. It turns out that gravitational energy cannot be defined in general. Gravity is *not* an interaction and does *not* have an associated energy.**

Hilbert's action – how do things fall

When Einstein discussed his work with David Hilbert, Hilbert found a way to do in a few weeks what had taken years for Einstein. Hilbert understood that general relativity *in empty space* could be described by an action integral, like all other physical systems.

Thus Hilbert set out to find the measure of change, as this is what an action describes, for motion due to gravity. Obviously, the measure must be observer invariant; in particular, it must include all possible changes of viewpoints, i.e. all the symmetries just described.

Motion due to gravity is determined by curvature. The only curvature measure independent of the observer is the Ricci scalar R and the cosmological constant Λ. It thus makes sense to expect that the change of space-time is described by an action S given by

$$S = \frac{c^3}{16\pi G} \int (R + 2\Lambda)\, dV . \tag{307}$$

The cosmological constant Λ (added some years later) appears as a mathematical possibility to describe the most general action that is diffeomorphism invariant. We will see below that its value in nature, though small, seems to be different from zero.

* Even though general relativity expressly forbids the existence of point particles, the approximation is useful in cases when the particle distances are large compared to their own size.

** In certain special circumstances, such as weak fields, slow motion, or an asymptotically flat space-time, we *can* define the integral over the G^{oo} component of the Einstein tensor as negative gravitational energy. Gravitational energy is thus only defined *approximately*, and only for our everyday life environment. Nevertheless, this approximation leads to the famous speculation that the total energy of the universe is zero. Do you agree? Page 436 Challenge 780 ny

The Hilbert action of a chunk of space-time is thus the integral of the Ricci scalar plus twice the cosmological constant over that chunk. The principle of least action states that space-time moves in such a way that this integral changes as little as possible.

– CS – to be finished – CS –

In summary, the question 'how do things move?' is answered by general relativity in the same way as by special relativity: *things follow the path of maximal aging*.

The symmetries of general relativity

The main symmetry of the Lagrangian of general relativity is called *diffeomorphism invariance*.

– CS – to be written – CS –

The field equations for empty space-time also show *scale symmetry*. This is the invariance of the equations after multiplication of all coordinates by a common numerical factor. In 1993, Torre and Anderson have shown that diffeomorphism symmetry and Ref. 354 trivial scale symmetry are the *only* symmetries of the vacuum field equations.

Apart from diffeomorphism symmetry, full general relativity, including mass–energy, has an additional symmetry which is not yet fully elucidated. Indeed, a complex symmetry connects the various possible initial conditions of the field equations; this symmetry is extremely complex and still is a topic of research. These fascinating investigations should give new insights into the classical description of the big bang.

Einstein's field equations

> [Einstein's general theory of relativity] cloaked the ghastly appearance of atheism.
> A witch hunter from Boston, around 1935

> Do you believe in god? Prepaid reply 50 words.
> Subsequent telegram by another to his hero Albert Einstein

> I believe in Spinoza's god, who reveals himself in the orderly harmony of what exists, not in a god who concerns himself with fates and actions of human beings.
> Albert Einstein's answer

The basis of many religious worries were Einstein's famous field equations. They contain Page 323 the full description of general relativity. As explained above, they follow from the maximum force in nature – or equivalently, from Hilbert's action – and are given by

$$G_{ab} = -\kappa\, T_{ab}$$
or
$$R_{ab} - \frac{1}{2} g_{ab} R - \Lambda g_{ab} = -\kappa\, T^{ab} \,. \qquad (308)$$

The constant κ, called the *gravitational coupling constant*, has been measured to be

$$\kappa = \frac{8\pi G}{c^4} = 2.1 \cdot 10^{-43} \,/\mathrm{N} \qquad (309)$$

and its small value – 2π divided by the maximum force $c^4/4G$ – reflects the weakness of gravity in everyday life, or better, the difficulty to bend space-time. The constant Λ, the so-called *cosmological constant*, corresponds to a vacuum energy volume density, or Page 424 pressure Λ/κ. Its low value is quite hard to measure. The presently favoured value is

$$\Lambda \approx 10^{-52} \,/\mathrm{m}^2 \quad \text{or} \quad \Lambda/\kappa \approx 0.5\,\mathrm{nJ/m^3} = 0.5\,\mathrm{nPa} \,. \qquad (310)$$

Ref. 355 Present measurements and simulations agree that this constant, even though it is numerically near to the square of the present radius of the universe, is a constant of nature that does not vary with time.

In summary, the field equations state that the curvature at a point is equal to the flow of energy–momentum through that point, taking into account the vacuum energy density. In other words, *energy–momentum tells space-time how to curve.**

The field equations of general relativity can be simplified for the case that speeds are

* Einstein arrived at his field equations using a number of intellectual guidelines that are called *principles* in the literature. Today, many of them are not seen as central any more. Nevertheless, we give a short overview.

- *Principle of general relativity*: all observers are equivalent; this principle, even though often stated, following the latest studies, is probably empty of any physical content.

- *Principle of general covariance*: the equations of physics must be stated in tensorial form; even though Ref. 356 it is known today that all equations can be written with tensors, even universal gravity, in many cases they require unphysical 'absolute' elements, i.e. quantities which affect others but are not affected themselves. This Page 628 unphysical idea is in contrast with the idea of *inter*action, as explained above.

- *Principle of minimal coupling*: the field equations of gravity are found from those of special relativity by taking the simplest possible generalization. Of course, now that the equations are known and tested experimentally, this principle is only of historical interest.

- *Equivalence principle*: acceleration is locally indistinguishable from gravitation; we used it to argue that space-time is semi-Riemannian, and that gravity is its curvature.

- *Mach's principle*: inertia is due to the interaction with the rest of the universe; this principle is correct, even though it is often maintained that it is not fulfilled in general relativity. In any case, it is not the essence Page 438 of general relativity.

- *Identity of gravitational and inertial mass*: this is included into the definition of mass from the outset, but restated ad infinitum in general relativity texts; it is implicitly used in the definition of the Riemann tensor.

- *Correspondence principle*: a new, more general theory, such as general relativity, must reduce to the previous theory, in this case universal gravity or special relativity, when restricted to the domains in which those are valid.

small. In that case $T_{oo} = \rho c^2$ and all other components of T vanish. Using the definition
Challenge 781 ny of the constant κ and setting $\varphi = (c^2/2)h_{oo}$ in $g_{ab} = \eta_{ab} + h_{ab}$, we find

$$\nabla^2\varphi = 4\pi\rho \quad \text{and} \quad \frac{d^2x}{dt^2} = -\nabla\varphi \tag{311}$$

which we know well, since it can be restated as follows: a body of mass m near a body of mass M is accelerated by

$$a = G\,\frac{M}{r^2}, \tag{312}$$

a value which is independent of the mass m of the falling body. And indeed, as noted already by Galileo, all bodies fall with the same acceleration, independently of their size, their mass, their colour, etc. Also in general relativity, gravitation is completely democratic.* The independence of free fall from the mass of the falling body is a result of describing space-time as a bent mattress. The masses moving on a mattress also move in the same way for all masses.

To get a feeling for the complete field equations, we have a short walk through their main properties. First of all, all motion due to space-time curvature is *reversible*, *differentiable* and thus *deterministic*. Note that only the complete motion, of space-time *and* matter *and* energy, has these properties. For particle motion only, motion is in fact *irreversible*, as in most examples of motion, some gravitational radiation is emitted.

Challenge 782 e

By contracting the field equations we find, for vanishing cosmological constant, the following expression for the Ricci scalar:

$$R = -\kappa T \ . \tag{317}$$

This result also implies the relation between the excess radius and the mass inside a sphere.

Challenge 783 ny

The field equations are *nonlinear* in the metric g, meaning that sums of solutions are

* Here is another way to show that general relativity fits with universal gravity. From the definition of the Riemann tensor we know that relative acceleration b_a and speed of nearby particles are related by

$$\nabla_e b_a = R_{ceda} v^c v^d \tag{313}$$

From the symmetries of R we know there is a φ such that $b_a = -\nabla_a \varphi$. That means that

$$\nabla_e b^a = \nabla_e \nabla^a \varphi = R^a_{ced} v^c v^d \tag{314}$$

which implies that

$$\begin{aligned}\Delta\varphi = \nabla_a \nabla^a \varphi &= R^a_{cad} v^c v^d \\ &= R_{cd} v^c v^d \\ &= \kappa(T_{cd} v^c v^d - T/2)\end{aligned} \tag{315}$$

Introducing $T_{ab} = \rho v_a v_b$ we get

$$\Delta\varphi = 4\pi G\rho \tag{316}$$

as we wanted to show.

not solutions. That makes the search for solutions rather difficult. For a complete solution of the field equations, initial and boundary conditions should be specified. The ways to do this form a specialized part of mathematical physics; it is not explored here.*

Albert Einstein used to say that the general relativity only provides the understanding of one side of the field equations (308), but not of the other. Can you see which side he meant?

Challenge 784 ny

What can we do of interest with these equations? In fact, to be honest, not much that we have not done already. Very few processes require the use of the full equations. Many textbooks on relativity even stop after writing them down! However, studying them is worthwhile. For example, one can show that the Schwarzschild solution is the *only* spherically symmetric solution. Similarly, in 1923, Birkhoff showed that every rotationally symmetric vacuum solution is static. This is the case even if masses themselves move, as for example during the collapse of a star.

Maybe the most beautiful application of the field equations are the various *movies* made of relativistic processes. The world wide web provides several of them; they allow one to see what happens when two black holes collide, what happens when an observer falls into a black hole, etc. For these movies, the field equations usually need to be solved directly, without approximations.*

– CS – more to be added – CS –

Another topic concerns *gravitational waves*. The full field equations show that waves are not harmonic, but nonlinear. Sine waves exist only approximately, for small amplitudes. Even more interestingly, if two waves collide, in many cases *singularities* are predicted to appear. The whole theme is still a research topic and might provide new insights for the quantization of general relativity in the coming years.

We end this section with a side note. Usually, the field equations are read in one sense only, as stating that energy–momentum produce curvature. One can also read them in the other way, calculating the energy–momentum needed to produce a given curvature. When this is done, one discovers that not all curved space-times are possible, as some would lead to *negative* energy (or mass) densities. Such solutions would contradict the mentioned limit on length-to-mass ratio for physical systems. The limit on length-to-mass ratios thus also restricts the range of possible curvatures of space-time.

More on the force limit

In case of a non-vanishing cosmological constant, the force limit holds exactly only if the constant Λ is positive; this is the case for the presently measured value, which is $\Lambda \approx 10^{-52}/\mathrm{m}^2$. Indeed, the radius–mass relation of black holes

Ref. 357
Ref. 288

$$2GM = Rc^2(1 - \frac{\Lambda}{3}R^2) \qquad (318)$$

* For more mathematical details, see the famous three-women-text in two volumes by Yvonne Choquet-Bruhat, Cecile DeWitt-Morette & Margaret Dillard-Bleick, *Analysis, Manifolds, and Physics*, North-Holland, 1996 and 2001, even though the first edition of this classic appeared in 1977.

* See for example, the http://math1.uibk.ac.at/~werner/black-Earth website.

implies that a radius-*independent* maximum force is valid only for positive or zero cosmological constant. For a negative cosmological constant the force limit would only be valid for infinitely small black holes. In the following, we take a pragmatic approach and note that a maximum force limit can be seen to imply a vanishing or positive cosmological constant. Obviously, the force limit does not specify the *value* of the constant; to achieve this, a second principle needs to be added. A straightforward formulation, using Page 341 the additional principle of a minimum force in nature, was proposed above.

One might ask also whether rotating or charged black holes change the argument that lead from maximum force to the derivation of general relativity. However, the derivation using the Raychaudhuri equation does not change. In fact, the only change of the argument appears with the inclusion of torsion, which changes the Raychaudhuri equation itself. As long as torsion plays no role, the derivation given above remains valid. The inclusion of torsion is still an open research issue.

Another question is how maximum force relates to scalar–tensor theories of gravity. If a particular scalar-tensor theory would obey the general horizon equation (209) then it would also show a maximum force. The general horizon equation must be obeyed both for *static* and for *dynamic* horizons. If that is the case, the specific scalar–tensor theory would be equivalent to general relativity, as it would allow one, using the argument of Jacobson, to deduce the usual field equations. This case can appear if the scalar field behaves like matter, i.e., if it has mass-energy like matter and curves space-time like matter. On the other hand, if in the particular scalar–tensor theory the general horizon equation (209) is not obeyed for *all moving* horizons – which is the general case, as scalar–tensor theories have more defining constants than general relativity – then the maximum force does not appear and the theory is not equivalent to general relativity. This connection also shows that an experimental test of the horizon equation for *static* horizons only is not sufficient to confirm general relativity; such a test rules out only some, but not all scalar–tensor theories.

Deducing universal gravity

In order to elevate the maximum force limit to a physical principle, it is not sufficient to show that it is a valid limit in nature. In addition, all properties of gravitation, including the full theory of general relativity, must be deduced from it. To make this argument more easy to follow, it is best split into several steps. First of all, we show that the force limit implies that in everyday life the inverse square law of universal gravity holds. Then we show that the main ideas of general relativity are included in the maximum force limit. Finally, we show that the full theory of general relativity follows. In other words, from this point onwards the force limit is assumed to be valid. We explore its consequences and compare them with the known properties of nature.

The result can be visualized also in another way. If the gravitational attraction between a central body and a satellite were *stronger* than it is, black holes would be smaller than they are; in that case the maximum force limit and the maximum speed could be exceeded. If on the other hand, gravitation were *weaker* than it is, a fast and accelerating observer would not be able to determine that the two bodies interact. In summary, a maximum force of $c^4/4G$ implies universal gravity. There is no difference in stating that all bodies attract through gravitation or in stating that there is a maximum force with the value

$c^4/4G$.

Deducing linearized general relativity

The next logical step is to show that a maximum force also implies general relativity. The naive approach is to repeat, step by step, the standard approach to general relativity.

Space-time curvature is a consequence of the fact that the speed of light is the maximum speed for *all* observers, even if they are located in a gravitational field. The gravitational red shift shows that in gravitational fields, clocks change their rate with height; that change, together with the constancy of the speed of light, implies space-time curvature. Gravity thus implies space-time curvature. The value of the curvature in the case of weak gravitational fields is completely fixed by the inverse square law of gravity. Since universal gravity follows from the maximum force, we deduce that maximum force implies space-time curvature.

Ref. 288

Apart from curvature, we must also check the other basic ideas of general relativity. The *principle of general relativity* states that all observers are equivalent; since the maximum force principle is stated to be valid for all observers, the principle of general relativity is contained in it. The *equivalence principle* states that locally, gravitation can be transformed away by changing to a suitable observer. This is also the case for the maximum force principle, which is claimed for all observers, thus also for observers that locally eliminate gravitation. *Mach's principle*, whose precise formulation varies most, states that only relative quantities should play a role in the description of nature. Since the maximum force is a relative quantity – in particular, the relation of mass and curvature remains – Mach's principle is also realized.

Free bodies in flat space move with constant speed. Using the equivalence principle this connection is changed to the statement that freely falling bodies move along geodesics. The maximum force principle keeps intact the statement that *space-time tells matter how to move*.

The curvature of space-time for weak gravitational fields is fixed by the inverse square law of gravity. Space curvature is thus present in the right amount around each mass. As Richard Feynman explains, by extending this result to all possible observers, all low curvature effects of gravitation follow. In particular, this implies the existence of linear (low-amplitude) gravitational waves and of the Lense–Thirring effect. Linearized general relativity thus follows from the maximum force principle.

Ref. 358

How to calculate the shape of geodesics

The other half of general relativity states that bodies fall along geodesics. All orbits are geodesics, thus curves with the longest proper time. It is thus useful to be able to calculate these trajectories.* To start, one needs to know the *shape of space-time*, that is the generalization of the shape of a two-dimensional surface. For a being living on the surface, it is usually described by the metric g_{ab}, which defines the distances between neighbouring points through

$$ds^2 = dx_a\, dx^a = g_{ab}(x)\, dx^a\, dx^b \,. \tag{319}$$

* This is a short section for the more curious; it can be skipped at first reading.

It is a famous exercise of calculus to show from this expression that a curve $x^a(s)$ depending on a well behaved (affine) parameter s is a timelike or spacelike (metric) *geodesic*, i.e. the longest possible path between the two events,* only if

Challenge 785 ny

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ad}\frac{\mathrm{d}x^d}{\mathrm{d}s}\right) = \frac{1}{2}\frac{\partial g_{bc}}{\partial x^a}\frac{\mathrm{d}x^b}{\mathrm{d}s}\frac{\mathrm{d}x^c}{\mathrm{d}s} , \tag{320}$$

as long as $\mathrm{d}s$ is different from zero along the path.* All bodies in free fall follow such geodesics. We showed above that the geodesic property implies that a stone thrown in the air falls back, except if it is thrown with a speed larger than the escape velocity. Expression (320) thus replaces both the expression $\mathrm{d}^2x/\mathrm{d}t^2 = -\nabla\varphi$ valid for falling bodies and the expression $\mathrm{d}^2x/\mathrm{d}t^2 = 0$ valid for freely floating bodies in special relativity.

Page 355

The path does not depend on the mass or on the material of the body. Therefore also *antimatter* falls along geodesics. In other words, antimatter and matter do not repel; they also attract each other. Interestingly, even experiments performed with normal matter can show this, if they are carefully evaluated. Are you able to find out why, using the details of the collision?

Ref. 359

Challenge 786 ny

For completeness, we mention that light follows *lightlike* or *null geodesics*, an affine parameter u exists, and the geodesics follow

$$\frac{\mathrm{d}^2x^a}{\mathrm{d}^2u} + \Gamma^a_{bc}\frac{\mathrm{d}x^b}{\mathrm{d}u}\frac{\mathrm{d}x^c}{\mathrm{d}u} = 0 \tag{324}$$

with the different condition

$$g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\mathrm{d}x^b}{\mathrm{d}u} = 0 . \tag{325}$$

Given all these definitions of various types of geodesics, what are the lines drawn in Figure 173 on page 350?

Challenge 787 ny

* We remember that in space in everyday life, geodesics are the shortest possible paths; however, in space-time in general relativity, geodesics are the longest possible paths. In both cases, they are the 'straightest' possible paths.

* This is often written as

$$\frac{\mathrm{d}^2x^a}{\mathrm{d}^2s} + \Gamma^a_{bc}\frac{\mathrm{d}x^b}{\mathrm{d}s}\frac{\mathrm{d}x^c}{\mathrm{d}s} = 0 \tag{321}$$

where the condition

$$g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}s}\frac{\mathrm{d}x^b}{\mathrm{d}s} = 1 \tag{322}$$

must be fulfilled, thus simply requiring that all the tangent vectors are *unit* vectors, and that $\mathrm{d}s \neq 0$ all along the path. The symbols Γ appearing above turn out to be defined as

$$\Gamma^a_{bc} = \left\{ \begin{matrix} a \\ bc \end{matrix} \right\} = \frac{1}{2}g^{ad}(\partial_b g_{dc} + \partial_c g_{db} - \partial_d g_{bc}) , \tag{323}$$

and are called *Christoffel symbols of the second kind* or simply the *metric connection*.

Mass in general relativity

The diffeomorphism invariance of general relativity makes life quite interesting. We will see that it allows us to say that we live on the *inside* of a hollow sphere, and that it does not allow us to say where energy actually is located. If energy cannot be located, what about mass? It soon became clear that mass, like energy, can be localized *only* if distant space-time is known to be flat. It is then possible to define a localized mass value by specifying with precision an intuitive idea: the mass is measured by the time a probe takes to orbit the unknown body.*

Challenge 788 ny
The intuitive mass definition *requires* flat space-time at infinity; it cannot be extended to other situations. In short, mass can only be localized if total mass can be defined. And *total mass* is defined only for asymptotically flat space-time. The only other notion of mass that is precise in general relativity is the *local mass density* at a point. In contrast, defining the mass contained in a region larger than a point but smaller than the entire space-time is not at all well understood.

Now that we can go on talking about mass without (too much) a bad conscience, we turn to the equations of motion.

Is gravity an interaction?

We tend to answer affirmatively, as in Galilean physics gravity was seen as an influence on the motion of bodies. In fact, we described it by a potential, implying that gravity *produces* motion. But let us be careful. A force or an interaction is what changes the motion of objects. However, we just saw that when two bodies attract each other through gravitation, both always remain in free fall. For example, the Moon circles the Earth because it continuously falls around it. Since any freely falling observer continuously remains at rest, the statement that gravity changes the motion of bodies is not correct for all observers. Indeed, we will soon discover that in a sense to be discussed shortly, the Moon and the Earth both follow 'straight' paths.

Is this correction of our idea of gravity only a question of words? Not at all. Since gravity is not an interaction, it is *not* due to a field and there is *no* potential.

Page 628
Let us check this strange result in yet another way. The most fundamental definition of 'interaction' is the difference between the whole and the sum of its parts. In the case of gravity, an observer in free fall could claim that nothing special is going on, independently of whether the other body is present or not, and could claim that gravity is not an interaction.

* This definition was formalized by Arnowitt, Deser and Misner, and since then is often called the ADM *mass*. The idea is to use the metric g_{ij} and to take the integral

$$m = \frac{1}{16\pi} \int_{S_R} (g_{ij,i}\nu_j - g_{ii,j}\nu_j)\, dA \qquad (326)$$

where S_R is the coordinate sphere of radius R, ν is the unit vector normal to the sphere and dA is the area element on the sphere. The limit exists if space-time is asymptotically flat and if the mass distribution is sufficiently concentrated. Mathematical physicists have also shown that for any manifold whose metric changes at infinity as Ref. 360

$$g_{ij} = (1 + f/r + O(1/r^2))\delta_{ij} \qquad (327)$$

the total mass is given by $M = 2f$.

However, that is going too far. An interaction transports energy between systems. We indeed found out that gravity can be said to transport energy only approximately. Gravitation is thus an interaction only approximately. But that is a sufficient reason to keep this characterization. In agreement with the strange conclusion, the concept of energy is not useful for gravity outside of everyday life. For the general case, namely for a general observer, gravity is thus fundamentally different from electricity or magnetism.

Another way to look at the issue is the following. Take a satellite orbiting Jupiter with energy–momentum $\mathbf{p} = m\mathbf{u}$. If we calculate the energy–momentum change along its path

Challenge 789 ny s, we get

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}s} = m\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}s} = m(\mathbf{e}_a\frac{\mathrm{d}\mathbf{u}^a}{\mathrm{d}s} + \frac{\mathrm{d}\mathbf{e}_a}{\mathrm{d}s}\mathbf{u}^a) = m\mathbf{e}_a(\frac{\mathrm{d}\mathbf{u}^a}{\mathrm{d}s} + \Gamma^a_{bd}\mathbf{u}^b\mathbf{u}^c) = 0 \qquad (328)$$

where $\mathbf{e}$ describes the unit vector along a coordinate axis. The energy–momentum change
Challenge 790 ny vanishes along any geodesic, as you might check. Therefore, the energy–momentum of this motion is conserved. In other words, *no* force is acting on the satellite. One could
Ref. 361 reply that in equation (328) the second term alone is the gravitational force. But the term
Challenge 791 ny can be made to vanish identically along any given world line. In short, nothing changes between two bodies in free fall around each other: gravity could be said not to be an
Challenge 792 n interaction. The properties of energy confirm this argument.

Of course, the conclusion that gravity is not an interaction is somewhat academic, as it contradicts daily life. But we will need it for the full understanding of motion later on. The behaviour of radiation confirms the deduction. In vacuum, radiation is always moving freely. In a sense, we can say that radiation always is in free fall. Strangely, since we called free fall the same as rest, we should conclude that radiation always is at rest. This is not wrong! We already saw that light cannot be accelerated.* We even saw that gravitational bending is not an acceleration, since light follows straight paths in space-time in this case as well. Even though light seems to slow down near masses for far away observers, it always moves at the speed of light locally. In short, even gravitation doesn't manage to move light.

There is another way to show that light is always at rest. A clock for an observer trying to reach the speed of light goes slower and slower. For light, in a sense, time stops: if one prefers, *light does not move*.

The essence of general relativity

If a maximum power or force appearing on horizons is the basis for relativity, one can ask whether physical systems other than space-time can also be described in this way. For special relativity, we found that all its main effects – such as a limit speed, Lorentz contraction or energy mass equivalence – are also found for dislocations in solids.

* Refraction, the slowdown of light inside matter, is not a counterexample. Strictly speaking, light inside matter is constantly being absorbed and re-emitted. In between these processes, light still propagates with the speed of light in vacuum. The whole process only *looks* like a slowdown in the macroscopic limit. The same applies to diffraction and to reflection. A list of apparent ways to bend light can be found on page 524; details of the quantum mechanical processes at their basis can be found on page 678.

Do systems analogous to general relativity exist? The historical answer is only partially positive. Several equations of general relativity are applicable to deformations of solids; since general relativity describes the deformation of the space-time mattress, many of its ideas are also applicable to the deformation behaviour of solids. Kröner has studied this analogy in great detail. Other systems with horizons, and thus with observables analogous to curvature, are found in certain liquids – where vortices play the role of black holes – and in certain quantum fluids for the propagation of light. Exploring such systems has become a research topic on its own. A full analogy of general relativity in a macroscopic system is available only since a few years. The analogy will be presented in the third part of our adventure; we will need an additional ingredient that is not visible at this point of our escalation.

Ref. 362
Ref. 364

Riemann gymnastics

Most books introduce curvature the hard way, namely historically,* using the Riemann curvature tensor. This is a short summary, so that you can understand that old stuff when you get it in your hands.

Above we saw that curvature is best described by a tensor. In 4 dimensions, this curvature tensor, usually called R, must be a quantity which allows us to calculate, among others, the area for any orientation of a 2-disk in space-time. Now, in four dimensions, orientations of a disk are defined with *two* 4-vectors; let us call them **p** and **q**. And instead of a disk, we take the *parallelogram* spanned by **p** and **q**. There are several possible definitions.

Challenge 793 e

The *Riemann-Christoffel curvature tensor* R is then defined as a quantity allowing to calculate the curvature $K(\mathbf{p},\mathbf{q})$ for the surface spanned by **p** and **q**, with area A, through

$$K(\mathbf{p},\mathbf{q}) = \frac{R\,\mathbf{pqpq}}{A^2(\mathbf{p},\mathbf{q})} = \frac{R_{abcd}\,p^a q^b p^c q^d}{(g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\gamma}g_{\beta\delta})p^\alpha q^\beta p^\gamma q^\delta} \qquad (329)$$

where, as usual, Latin indices a, b, c, d, etc. run from 0 to 3, as do Greek indices here, and a *summation* is implied when an index name appears twice. Obviously R is a tensor, of rank 4. This tensor thus describes the *intrinsic* curvature of a space-time only. In contrast, the metric g describes the complete *shape* of the surface, not only the curvature. The curvature is thus the physical quantity of relevance locally, and physical descriptions therefore use only the *Riemann*** *tensor* R or quantities derived from it.***

* This is a short section for the more curious; it can be skipped at first reading.

** Bernhard Riemann (b. 1826 Breselenz, d. 1866 Selasca), important German mathematician.

*** Above, we showed that space-time is curved by noting changes in clock rates, in meter bar lengths and in light propagation. Such experiments most easily provide the metric g. We know that space-time is described by a four-dimensional manifold M with a metric g_{ab} that locally, at each space-time point, is a Minkowski metric with all its properties. Such a manifold is called a *Riemannian manifold*. Only such a metric allows one to define a local inertial system, i.e. a local Minkowski space-time at every space-time point. In particular, we have

$$g_{ab} = 1/g^{ab} \quad \text{and} \quad g_a{}^b = g^a{}_b = \delta^a_b\,. \qquad (330)$$

How are curvature and metric related? The solution usually occupies a large number of pages in relativity

But we can forget the just mentioned definition of curvature. There is a second, more physical way to look at the Riemann tensor. We know that curvature means gravity. As said above, gravity means that when two nearby particles move freely with the same velocity and the same direction, the distance between these two particles changes. In other words, the local effect of gravity is *relative acceleration* of nearby particles.

Challenge 794 e

It turns out that the tensor R describes precisely this relative acceleration, i.e. what we called the *tidal effects* earlier on. Obviously, the relative acceleration $\mathbf{b}$ increases with the separation $\mathbf{d}$ and the square (why?) of the speed $\mathbf{u}$ of the two particles. Therefore we can also define R as a (generalized) proportionality factor among these quantities:

Challenge 795 ny

$$\mathbf{b} = R\,\mathbf{u}\,\mathbf{u}\,\mathbf{d} \quad \text{or, more clearly} \quad b^a = R^a{}_{bcd}\, u^b\, u^c\, d^d\ . \tag{333}$$

The components of the Riemann curvature tensor have the dimension of an inverse square length. Since it contains all information about intrinsic curvature, we conclude that if R vanishes in a region, space-time in that region is flat. This connection is easily deduced from this second definition.*

Challenge 796 ny

A final way to define the tensor R is the following. For a free falling observer, the metric g_{ab} is given by the metric η_{ab} from special relativity. In its neighbourhood, we have

$$\begin{aligned} g_{ab} &= \eta_{ab} + \frac{1}{3} R_{acbd} x^c x^d + O(x^3) \\ &= \frac{1}{2}(\partial_c \partial_d g_{ab}) x^c x^d + O(x^3)\ . \end{aligned} \tag{335}$$

The curvature term thus describes the dependence of the space-time metric from flat space-time. The curvature tensor R is a large beast; it has $4^4 = 256$ components at each point of space-time; however, its symmetry properties reduce them to twenty independ-

books; just for information, the relation is

$$R^a{}_{bcd} = \frac{\partial \Gamma^a{}_{bd}}{\partial x^c} - \frac{\partial \Gamma^a{}_{bc}}{\partial x^d} + \Gamma^a{}_{ec}\Gamma^e{}_{bd} - \Gamma^a{}_{fd}\Gamma^f{}_{bc}\ . \tag{331}$$

The curvature tensor is built from the second derivatives of the metric. On the other hand, we can also determine the metric if the curvature is known, using

$$g = ...R... \tag{332}$$

In other words, either the Riemann tensor R or the metric g specify the whole situation of a space-time.

* This second definition is also called the definition through *geodesic deviation*. It is of course not evident that it coincides with the first. For an explicit proof, see the literature. There is also a third way to picture the tensor R, a more mathematical one, namely the original way Riemann introduced it. If one parallel transports a vector $\mathbf{w}$ around a parallelogram formed by two vectors $\mathbf{u}$ and $\mathbf{v}$, each of length ε, the vector $\mathbf{w}$ is changed to $\mathbf{w} + \delta\mathbf{w}$. One then has

Ref. 363

$$\delta\mathbf{w} = -\varepsilon^2 R\,\mathbf{u}\,\mathbf{v}\,\mathbf{w} + \quad \text{higher order terms} \quad . \tag{334}$$

More about the geodesic deviation can be found out by studying the behaviour of the famous south-pointing carriage. This device, common in China before the compass was discovered, only works if the world is flat. Indeed, on a curved surface, after following a large closed path, it will show a different direction than at the start of the trip. Can you explain why?

Page 159
Challenge 797 ny

ent numbers.* The actual number of importance in physical problems is still smaller, namely only ten. These are the components of the Ricci tensor, which can be defined with the help of the Riemann tensor by contraction, i.e. by setting

$$R_{bc} = R^{a}{}_{bac} \; . \tag{338}$$

Its components, like those of the Riemann tensor, are inverse square lengths.

Challenge 799 ny Can you confirm the relation $R_{abcd}R^{abcd} = 48m^2/r^6$ for the Schwarzschild solution:?

10. Why can we see the stars? – Motion in the universe

> Zwei Dinge erfüllen das Gemüt mit immer neuer und zunehmender Bewunderung und Ehrfurcht, je öfter und anhaltender sich das Nachdenken damit beschäftigt: der bestirnte Himmel über mir und das moralische Gesetz in mir.*
>
> Immanuel Kant (1724–1804)

On clear nights, between two and five thousand stars are visible with the naked eye. Several hundreds of them have names. Indeed, in all parts of the world, the stars and the constellations they form are seen as memories of ancient events, and stories are told about them.** But the simple fact that we can *see* the stars is the basis for a story much more fantastic than all myths. It touches almost all aspects of modern physics.

Which stars do we see?

> Democritus says [about the milky way] that it is a region of light emanating from numerous stars small and near to each other, of which the grouping produces the brightness of the whole.
>
> Ref. 366 Aetius, *Opinions*.

The stars we see on a clear night are mainly the brightest of our nearest neighbours in the surrounding region of the milky way. They lie at distances between four and a few

---

* The second definition indeed shows that the Riemann tensor is symmetric in certain indices and antisymmetric in others:

Challenge 798 ny

$$R_{abcd} = R_{cdab} \quad , \quad R_{abcd} = -R_{bacd} = -R_{abdc} \tag{336}$$

which also imply that many components vanish. Of importance is also the relation

$$R_{abcd} + R_{adbc} + R_{acdb} = 0 \; . \tag{337}$$

Note that the order of the indices depends on the book one uses, and is not standardized. The list of invariants which can be constructed from R is long. We mention that $\frac{1}{2}\varepsilon^{abcd}R_{cd}{}^{ef}R_{abef}$, namely the product ${}^*R\,R$ of the Riemann tensor with its dual, is the invariant characterizing the Lense–Thirring effect.

* Two things fill the mind with ever new and increasing admiration and awe, the more often and persistently thought considers them: the starred sky above me and the moral law inside me. Ref. 365

** About the myths around the stars and the constellations, see e.g. the text by G. Fasching, *Sternbilder und ihre Mythen*, Springer Verlag, 1993. On the internet there are also the beautiful http://www.astro.wisc.edu/~dolan/constellations/constellations.html and http://www.astro.uiuc.edu/~kaler/sow/sow.html websites.

Figure 189 The Andromeda nebula M31, our neighbour galaxy (and the 31st member of the Messier object listing)

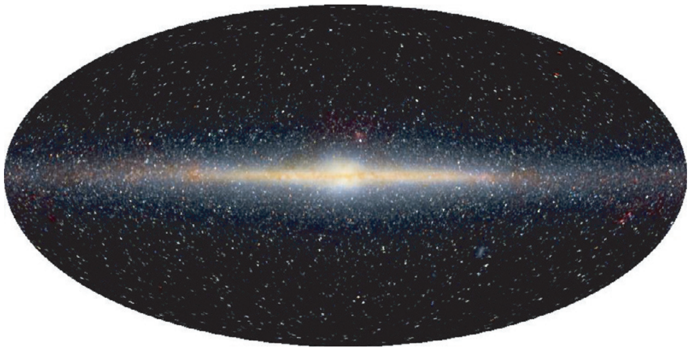


Figure 190 How our galaxy looks in the infrared

thousand light years from us. Roughly speaking, in our environment there is a star about every 400 cubic light years.

Almost all visible stars are from our own galaxy. The only extragalactic object *constantly* visible to the naked eye in the northern hemisphere is the so-called Andromeda nebula, shown enlarged in Figure 189. It is a whole galaxy like our own, as Immanuel Kant already had conjectured in 1755. Several extragalactic objects are visible with the naked eye in the southern hemisphere: the Tarantula Nebula, as well as the large and the small Magellanic cloud. The Magellanic clouds are neighbour galaxies to our own. Other, temporary exceptions are the rare *novae*, exploding stars which can be seen also if they appear in nearby galaxies, or the still rarer *supernovae*, which can often be seen even in faraway galaxies.

In fact, the visible stars are special also in other respects. For example, telescopes show that about half of them are in fact double; they consist of two stars circling around each other, as in the case of Sirius. Measuring the orbits they follow around each other allows one to determine their masses. Can you explain how? Challenge 800 ny

Is the universe different from our milky way? Yes, it is. There are several arguments. First of all, our galaxy – that is just the Greek original of the term 'milky way' – is *flattened*,

Figure 191 The elliptical galaxy NGC 205 (the 205^{th} member of the New Galactic Catalogue)

due to its rotation. If the galaxy rotates, there must be other masses which determine the background with respect to which this rotation takes place. In fact, there is a huge number of other galaxies – about 10^{11} – in the universe, a discovery dating only from the twentieth century.

Why did this happen so late? Well, people had the same difficulty as when the shape of the Earth had to be determined. They had to understand that the galaxy is not only a milky strip seen in clear nights, but an actual physical system, made of about 10^{11} stars gravitating around each other.* As in the case of the Earth, the galaxy was found to have a three-dimensional *shape*; it is shown in Figure 190. Our galaxy is a flat and circular structure, with a diameter of 100 000 light years; in the centre, it has a spherical bulge. As said before, it rotates once in about 200 to 250 million years. (Can you guess how this is measured?) The rotation is quite slow: since the Sun was formed, it has made only about 20 to 25 full turns around the centre.

Challenge 801 ny

It is even possible to measure the *mass* of our galaxy. The trick is to use a binary pulsar on its outskirts. If it is observed for many years, one can deduce its acceleration around the galactic centre, as the pulsar reacts with a frequency shift which can be measured on Earth. However, many decades of observation are needed and many spurious effects have to be eliminated. Nevertheless, such measurements are ongoing. Present estimates put the mass of our galaxy at $10^{41\pm1}$ kg.

Ref. 367

* The milky way, or *galaxy* in Greek, was said to have originated when , the main Greek god, tried to let his son feed at Hera's breast in order to make him immortal; the young Heracles, in a sign showing his future strength, sucked so forcefully that the milk splashed all over the sky.

Figure 192 The colliding galaxies M51 and M110

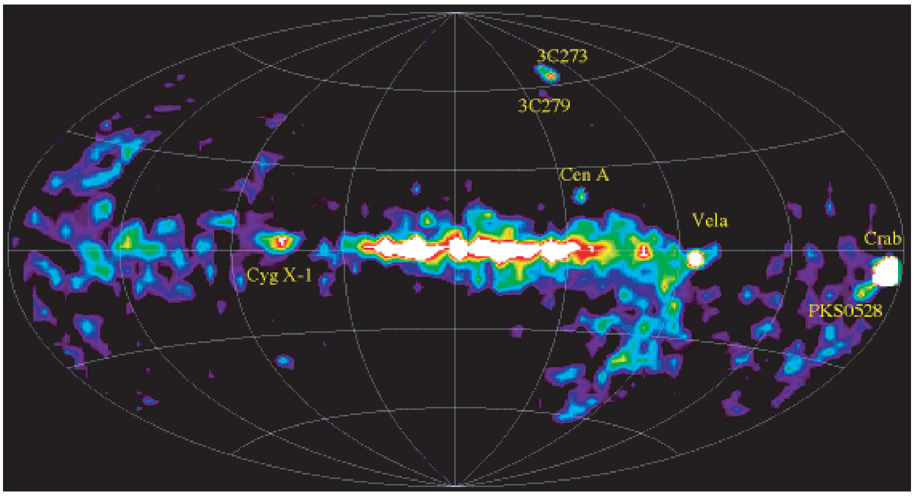


Figure 193 The X-rays in the night sky, between 1 and 30 MeV

What do we see at night?

Astrophysics leads to a strange conclusion about matter, quite different from what we are used to think in classical physics: *the matter observed in the sky is found in clouds*. *Clouds* are systems in which the matter density diminishes with the distance from the centre, with no clear border and with no clear size. Most astrophysical objects are best described as clouds.

The Earth is also a cloud, if we take its atmosphere, its magnetosphere and its dust ring around it as part of it. The Sun is a cloud. It is a gas ball first of all, but is even more a cloud if we take into consideration its protuberances, its heliosphere, the solar

Figure to be included

Figure 194 Rotating clouds emitting jets along the axis; top left: a composite image (visible and infrared) of the galaxy 0313-192 (Hubble Space Telescope), top right: the young star in formation DG Tauri B seen edge on, bottom left: a diagram showing a pulsar, bottom right: a diagram of an accreting black hole

wind it generates and its magnetosphere. The solar system is a cloud if we consider its comet cloud, its asteroid belt and its local interstellar gas cloud. The galaxy is a cloud if we remember its matter distribution and cloud of the cosmic radiation it is surrounded with. In fact, even people can be seen as clouds, as every person is surrounded by gases, little dust particles from skin, vapour, etc.

In the universe, *almost all of the clouds are plasma clouds*. A *plasma* is an ionized gas, such as fire, lightning, the inside of neon tubes, the Sun etc. At least 99.9 % of all matter in the universe is in the form of plasmas. Only an exceptionally small percentage exists in solid or liquid form, such as toasters, subways or their users.

Finally, clouds in the universe have certain common properties. Clouds seen in the universe, when undisturbed by collisions or other interactions from neighbouring objects, are rotating. Most clouds are therefore flattened. Finally, undisturbed rotating clouds usually emit something along the rotation axis. This basic structure has been observed for young stars, for pulsars, for galaxies, for quasars and for many other systems. Figure 194 gives an overview.

In summary, the universe is mostly made of rotating, flattened plasma clouds emitting jets along their axes. A more detailed overview of the information collected by modern Ref. 368 astronomy and astrophysics about various clouds in the universe is given in Table 35.*

Table 35 Some observations about the universe

Aspect	Main properties	Value
Phenomena		
galaxy formation	observed by Hubble	several times
	trigger event	unknown

* More details about the universe can be found in the beautiful text by W.J. Kaufmann & R.A. Friedman, *Universe*, fifth edition, W.H. Freeman & Co., 1999. The most recent discoveries are best followed on the http://sci.esa.int and http://hubble.nasa.gov websites.

Aspect	Main properties	Value
galactic collisions	momentum	$p \approx ...$
	star formation	
star formation	cloud collapse	
novae	new bright stars,	$L > ...$
	later surrounded by bubble	$R \approx t \cdot c/100$
supernovae	new bright star,	$L > ...$
	matter forms	
hypernovae, optical bursts		
gamma ray bursts	luminosity	up to $\cdot 10^{45}$ W, about one per cent of the whole visible universe
	energy	*c.* 10^{46} J
	duration	*c.* 0.015– 1000 s
	observed number	*c.* 2 per day
radio sources		
X-ray sources		
cosmic rays	energy	from 0 eV to 10^{22} eV
gravitational lensing	light bending	
comets	recurrence, evaporation	
meteorites	age	up to $4.6 \cdot 10^9$ a
Observed components		
intergalactic space	mass density	...
quasars	red-shift	up to 5.8
	luminosity	..., about the same as one galaxy
galaxy superclusters	number	*c.* 10^8 inside horizon
our own local supercluster		with about 4000 galaxies
galaxy groups		100 Zm, with a dozen up to 1000 galaxies
our local group		with 30 galaxies
galaxies	size	0.5 to 2 Zm
	number	*c.* 10^{11} inside horizon
	containing	10 to 400 globular clusters
	containing	typically 10^{11} stars each
	containing	typically one supermassive and several intermediate mass black holes
our galaxy	diameter	1.0(0.1) Zm
	mass	10^{42} kg or $5 \cdot 10^{11}$ solar masses Ref. 369
	containing	100 globular clusters each with 1 million stars
	speed	600 km/s towards Hydra-Centaurus

Aspect	Main properties	Value
globular clusters (e.g. M15)	containing	thousands of stars, an intermediate mass black hole
	age	max. $12 \cdot 10^9$ a, oldest known objects
nebulae, clouds	composition	dust, oxygen, hydrogen
our local interstellar cloud	size	20 light years
	composition	atomic hydrogen at 7500 K
star systems	types	orbiting double stars, over 70 stars orbited by brown dwarfs, several planetary systems
our solar system	size	2 light years (Oort cloud)
our solar system	speed	370 km/s from Aquarius towards Leo
stars	mass	up to 130 solar masses (except when stars fuse) Ref. 370
giants and supergiants	large size	up to 10^{12} m
brown dwarfs	low mass	below 0.072 solar masses
	low temperature	below 2800 K Ref. 371
L dwarfs	low temperature	
T dwarfs	low temperature	
white dwarfs	small radius	$r \approx 5\,000$ km
	high temperature	
neutron stars	nuclear mass density, small size	$\rho \approx 10^{17}$ kg/m^3, $r \approx 10$ km
jet sources		
central compact objects		
emitters of X-ray bursts	X-ray emission	
pulsars	periodic radio emission	
	mass	below around 25 solar masses
magnetars	high magnetic fields	up to 10^{11} T and more Ref. 372
(soft gamma repeaters, anomalous X-ray pulsars)		
	mass	above 25 solar masses Ref. 373
black holes	horizon radius	$r = 2GM/c^2$, observed mass range from 1 to 100 million solar masses
General properties		
cosmic horizon	distance	*c.* 10^{26} m=100 Ym
expansion	Hubble's constant	between 59 km s^{-1} Mpc^{-1} and 70 km s^{-1} Mpc^{-1}, or *c.* $2 \cdot 10^{-18}$ /s
	h	0.73(3)
'age' of the universe		13.7(2) Ga
large size shape	space curvature	almost vanishing, $k = \Omega k = 0$
vacuum energy density	finite	0.5 nJ/m^3 or $\Omega_\Lambda = 0.73$ for $k = 0$

Aspect	Main properties	Value
		no evidence for time dependence
large size shape	topology	simple in our galactic environment, unknown at large scales
dimensions	number	3 for space, 1 for time, at low and moderate energies
matter	density	2 to $11 \cdot 10^{-27}$ kg/m^3 or 1 to 6 hydrogen atoms per cubic metre
		Ωm = 0.25
baryons	density	$\Omega_b = 0.04$, one sixth of the previous (included in Ωm)
photons	number density	4 to $5 \cdot 10^8$ /m^3 = 1.7 to $2.1 \cdot 10^{-31}$ kg/m^3
	energy density	Ωr $= 4.6 \cdot 10^{-5}$
neutrinos	energy density	$\Omega\nu$ unclear, probably small
average temperature	photons	2.725(2) K
	matter	*c.* 0 K
	neutrinos	not measured, predicted value is 2 K
perturbations: photon anisotropy	$\Delta T/T$	$1 \cdot 10^{-5}$
perturbations: density amplitude		$A = 0.8(1)$
perturbations: spectral index		$n = 0.97(3)$
perturbations: tensor to scalar ratio		$r < 0.53$ with 95% confidence
ionization optical depth		$\tau = 0.15(7)$
decoupling		$z = 1100$

But since we are speaking of what we see in the sky, we need to clarify a general issue.

What is the universe?

> I'm astounded by people who want to 'know' the universe when it's hard enough to find your way around Chinatown.
>
> Woody Allen

The universe, as the name says, is what turns around us at night. For a physicist, at least three definitions are possible for the term 'universe':

The *(visible) universe* is the totality of all observable mass and energy. This includes everything inside the cosmological horizon. Since the horizon is moving away from us, the observable mass and energy is constantly increasing. The content of the term 'visible universe' is thus not fixed in time. (What is the origin of this increase? We will come back to this issue later on.)

Figure 195 The universe is full of galaxies – here the Perseus cluster

The *(believed) universe* is the totality of all mass and energy, *including* any parts of them which are not visible. Numerous books of general relativity state that there definitely exists matter or energy beyond the observation boundaries. We explain the origin of this belief below.

The *(full) universe* is the sum of matter, energy *as well as* space-time itself.

These definitions are often mixed up in physical and philosophical discussions. There is *no* generally accepted consensus on the terms, so one has to be careful. In this adventure, when we use the term 'universe', we imply the *last* definition only. We will discover repeatedly that without clear distinction between the definitions the complete ascent of Motion Mountain becomes impossible. (For example: Is the matter and energy in the full universe the same as the visible universe?) Challenge 802 ny

Note that the 'size' of the visible universe, or better, the distance to its horizon, is a quantity which *can* be imagined. The value of 10^{26} m is not beyond imagination. If one took all the iron from the Earth's core and made it into a wire reaching the universe, how thick would it be? The answer might surprise you. Also the content of the universe is clearly finite. There are about as many visible *galaxies* in the universe as there are grains in a cubic metre of sand. To expand on the comparison, can you deduce how much space you need to contain all the flour you would get if every little speck represented one star? Challenge 803 ny Challenge 804 ny

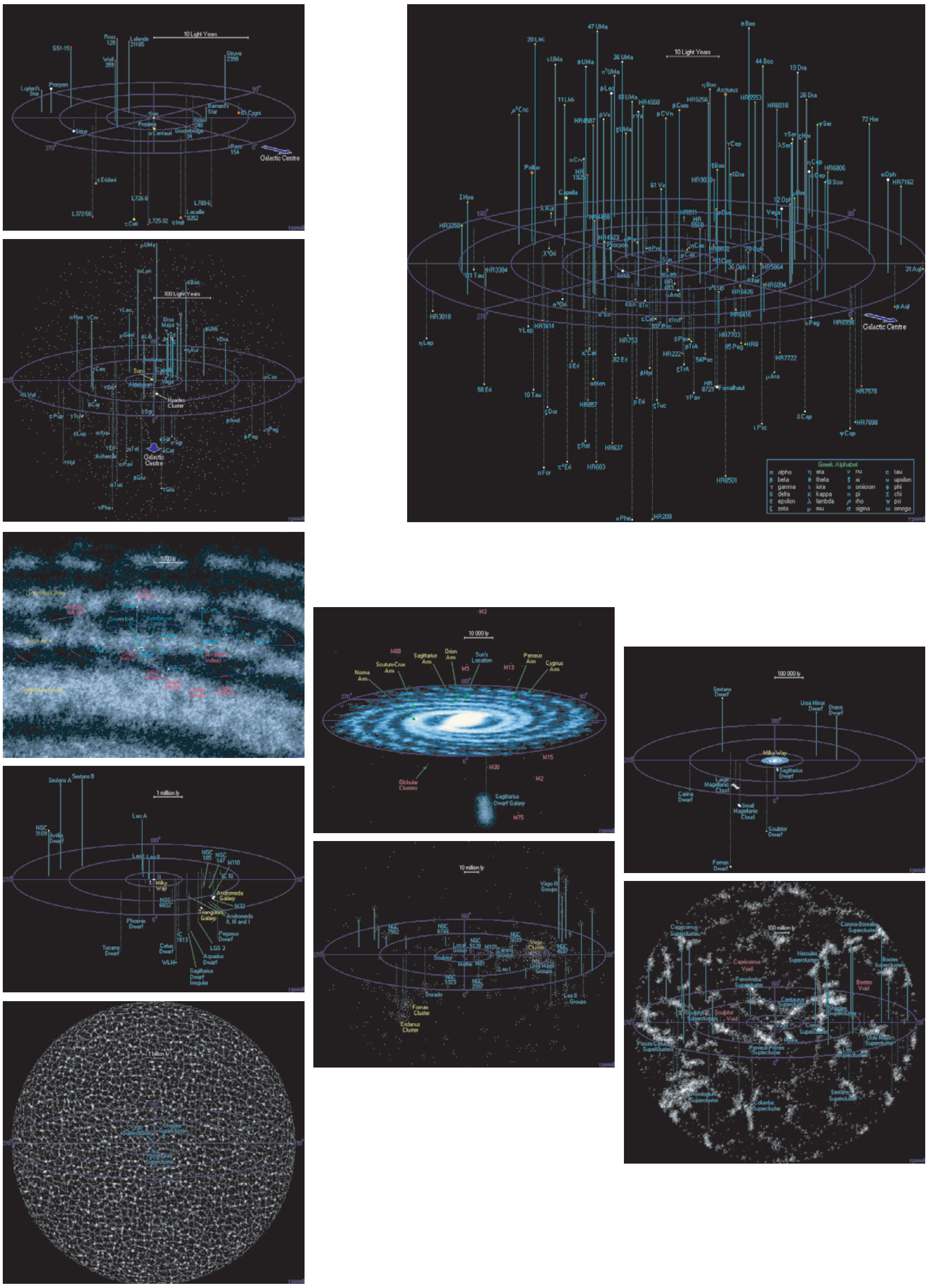


Figure 196 An atlas of our cosmic environment; illustrations up to 12.5, 50, 250, 5 000, 50 000, 500 000, 5 million, 100 million, 1 000 million and 14 000 million light years (© Richard Powell, http://www.anzwers.org/free/universe)

The colour and the motion of the stars

Ἠ τοι μὲν πρώτιστα Ξ΄αος γένετ΄ ... *
Hesiod, *Theogony*.

Obviously, the universe is full of motion. To get to know the universe a bit, it is useful to measure the speed and position of as many objects in it as possible. In the twentieth century, a large number of such observations have been performed on stars and galaxies. Challenge 805 ny (Can you imagine how distance and velocity are determined?) This wealth of data can be summed up in two points.

First of all, on large scales, i.e. averaged over about five hundred million light years, the matter density in the universe is *homogeneous* and *isotropic*. Obviously, at smaller scales inhomogeneities exist, such as galaxies or cheese cakes. Our galaxy for example is Ref. 374 neither isotropic nor homogeneous. But at large scales the differences average out. This large scale homogeneity of matter position is often called the *cosmological principle*.

The second point about the universe is even more important. In the 1920s, independ- Ref. 375 ently, Carl Wirtz, Knut Lundmark and Gustaf Stromberg showed that on the whole, galaxies move away from the Earth, and the more so, the more they were distant. There are a few exceptions for nearby galaxies, such as the Andromeda nebula itself; but in general, the speed of flight v of an object increases with distance d. In 1929, the US-American astronomer Edwin Hubble* published the first measurement of the relation between speed and distance. Despite his use of incorrect length scales he found a relation

$$v = H\,d\,, \tag{339}$$

where the proportionality constant H, so-called *Hubble constant*, is known today to have a value between $59\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. (Hubble's own value was so far outside this range that it is not cited any more.) For example, a star at a distance of 2 Mpc** is moving away from Earth with a speed between 118 km/s and 140 km/s, and proportionally more for stars further away.

In fact, the discovery by Wirtz and Lundmark implies that every galaxy moves away Challenge 806 ny from all the others. (Why?) In other words, the matter in the universe is *expanding*. The scale of this expansion and the enormous dimensions involved are amazing. The motion of all the thousand million galaxy groups in the sky is described by the single equation (339)! Of course, some deviations are observed for nearby galaxies, as mentioned above, and for far away galaxies, as we will see.

The cosmological principle and the expansion taken together imply that the universe cannot be older than that time when it was of vanishing size; the universe thus has a *finite age*. Including the evolution equations, as explained in more detail below, the Hubble

* 'Verily, at first chaos came to be ...' The Theogony, attributed to the probably mythical Hesiodos, was finalized around 700 BCE. It can be read in English and Greek on the http://perseus.csad.ox.ac.uk/cgi-bin/ptext?lookup=Hes.+Th.+5 web site. The famous citation is from verse 117.

* Edwin Powell Hubble (1889–1953), important US-American astronomer. After being an athlete and taking a law degree, he returned to his childhood passion of the stars; he finally proved Immanuel Kant's 1755 conjecture that the Andromeda nebula was a galaxy like our own. He thus showed that the milky way is only a tiny part of the universe.

Page 1071 ** A megaparsec or Mpc is a distance of 30.8 Zm.

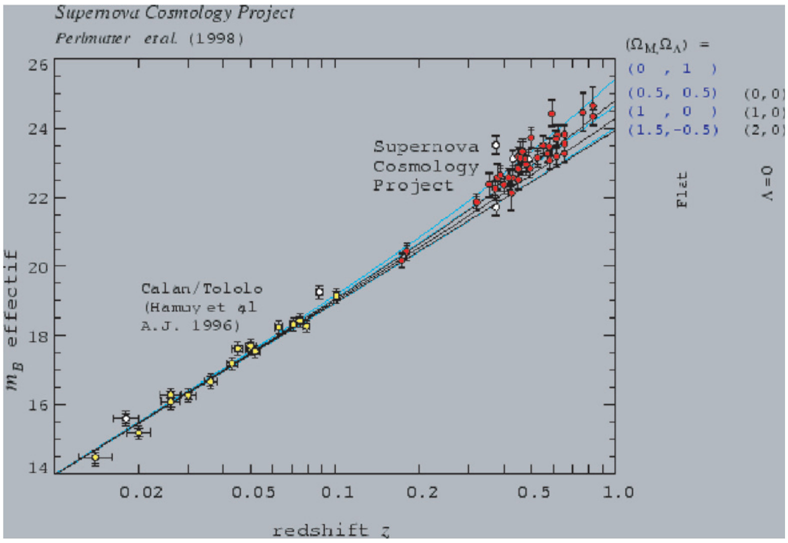


Figure 197 The relation between star distance and star velocity

constant points to an age value of around fourteen thousand million years, with an error of about a sixth of this value. That also means that the universe has a *horizon*, i.e. a finite maximum distance for sources whose signals can arrive on Earth. Signals from sources beyond the horizon cannot reach us.

Since the universe is expanding, in the past it has been much smaller and thus much denser than it is now. It turns out that it also has been hotter. George Gamow* predicted in 1948 that since hot objects radiate light, the sky cannot be completely black at night, but must be filled with black body radiation emitted during the times it was 'in heat'. That radiation, called the *background radiation*, must have cooled down due to the expansion of the universe. (Can you confirm this?) Despite various similar predictions by other authors, in one of the most famous cases of missed scientific communication, the radiation was found only much later, by two researchers completely unaware of all this work. A famous paper in 1964 by Doroshkevich and Novikov had even stated that the antenna used by the (unaware) later discoverers was the best device to search for the radiation! In any case, only in 1965, Arno Penzias and Robert Wilson discovered the radiation, in one of the most beautiful discoveries of physics, for which both later received the Nobel prize for physics. The radiation turns out to be described by the black body radiation for a body with a temperature of 2.7 K; it follows the black body dependence to the precision of about 1 part in 10^4.

Ref. 376 Challenge 807 ny Ref. 377 Ref. 378

But apart from expansion and cooling, the past fourteen thousand million years also produced a few other memorable events.

* George Gamow (b. 1904 Odessa , d. 1968 St. Boulder), Russian-American physicist; he explained alpha decay as a tunnelling effect and predicted the microwave background. He wrote the first successful popular science texts, such as *1, 2, 3, infinity* and the *Mr. Thompkins* series, which later were imitated by many others.

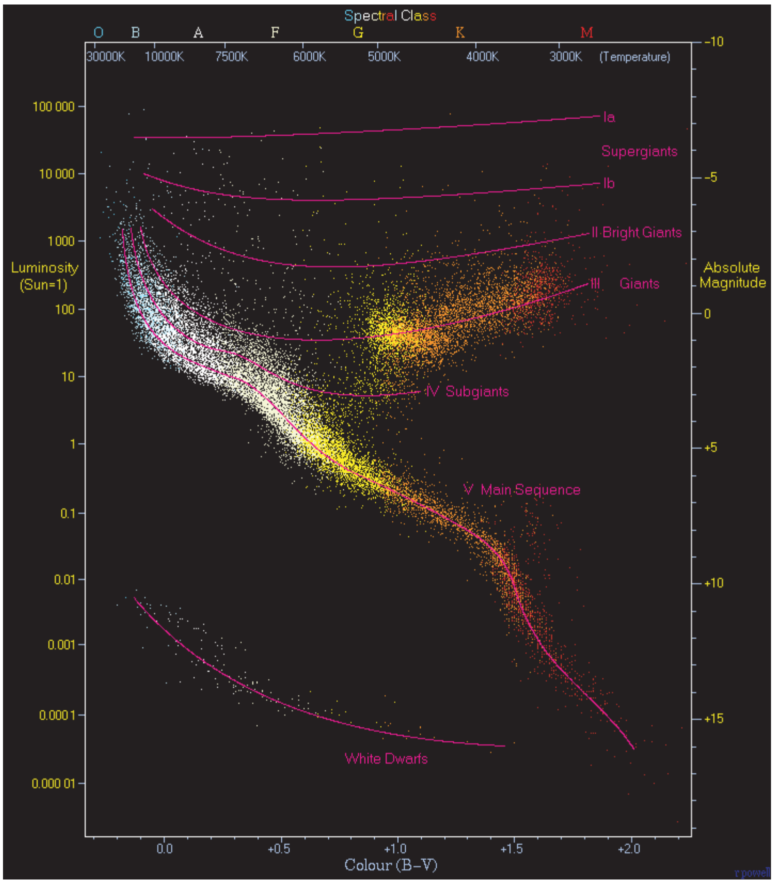


Figure 198 The Hertzsprung–Russell diagram (© Richard Powell)

Do stars shine every night?

Don't the stars shine beautifully? I am the only person in the world who knows why they do.
Friedrich (Fritz) Houtermans (1903–1966)

Stars seem to be there for ever. In fact, every now and then a new star appears in the sky: a *nova*. The name is Latin and means 'new'. Especially bright novae are called *supernovae*. Novae and similar phenomena remind us that stars usually live much longer than humans, but that like us, they are born and die.

It turns out that one can follow the age of a star in the so-called *Hertzsprung–Russell diagram*. The diagram, central to every book on astronomy, is a beautiful example of a standard method used by astrophysicists: collecting statistics over many examples of a type of object, one can deduce their life cycle, even though the lifetime is much longer than that of a human. For example, it is possible, by clever use of the diagram, to estimate the age of stellar clusters, and thus arrive at a minimum age of the universe. The result is around fourteen thousand million years.

One conclusion is basic: since stars shine, they also *die*. In other words, stars can be seen if they are born but not yet dead at the moment of light emission. That also leads to restrictions on their visibility, especially for high red-shifts. Indeed, the objects observed at large distances, such as quasars, are not stars, but much more massive and bright systems. These mechanisms are still being studied by astrophysicists.

On the other hand, since the stars shine, they were also *formed* somehow. The fascinating details of these investigations are part of astrophysics and will not be explored here.

Yet we do not have the full answer to our question. Why do stars shine at all? Clearly, they shine because they are hot. They are hot because of nuclear reactions in their interior. Page 858 We will discuss these processes in more detail in the chapter on the nucleus.

A short history of the universe

Ref. 379

> The soul is a spark of the substance of the stars.
> Heraclitus of Ephesus (*c.* 540 to *c.* 480 bce)

The adventures the universe has experienced, or better, the adventures the matter and radiation inside it have experienced, are summarized in Table 36.* The steps not yet discussed will be studied in quantum theory. This history table has applications no physicist would have imagined. The sequence is so beautiful and impressive that nowadays it is used in certain psychotherapies to point out to people the story behind their existence and to remind them of their own worth. Enjoy.

Table 36 A short history of the universe

Time from now[a]	Time from big bang[b]	Event	Temperature
$\approx 14 \cdot 10^9$ a	$\approx t_{\mathrm{Pl}}$ [b]	Time, space, matter and initial conditions indeterminate	10^{32} K $\approx T_{\mathrm{Pl}}$
$13 \cdot 10^9$ a	*c.* 800 t_{Pl} $\approx 10^{-42}$ s	Distinction of space-time from matter and radiation, initial conditions determinate	10^{30} K
	10^{-35} s to 10^{-32} s	**Inflation & GUT epoch** starts; strong and electroweak interactions diverge	$5 \cdot 10^{26}$ K
	10^{-12} s	Antiquarks annihilate; electromagnetic and weak interaction separate	10^{15} K
	$2 \cdot 10^{-6}$ s	Quarks get confined into hadrons; universe is a plasma	10^{13} K
		Positrons annihilate	
	0.3 s	Universe becomes transparent for neutrinos	10^{10} K
	a few seconds	**Nucleosynthesis:** D, ^{4}He, ^{3}He and ^{7}Li *nuclei* form; radiation still dominates	10^9 K
	2500 a	**Matter domination** starts; density perturbations magnify	75 000 K

* On the remote history of the universe, see the excellent texts by G. Börner, *The Early Universe – Facts & Fiction*, 3rd edition, Springer Verlag, 1993, or Barry Parker, *Creation – The Story of the Origin and the Evolution of the Universe*, Plenum Press, 1988. For an excellent popular text, see M. Longair, *Our Evolving Universe*, Cambridge University Press, 1996.

Time from now[a]	Time from big bang[b]	Event	Temperature
$z = 1100$	380 000 a	**Recombination**: during these latter stages of the big bang, H, He and Li *atoms* form, and the universe becomes 'transparent' for light, as matter and radiation decouple, i.e. as they acquire different temperatures; the 'night' sky starts to get darker and darker	3000 K
		Sky is almost black except for black body radiation	$T_\gamma = T_o(1+z)$
$z = 10 - 30$		**Galaxy formation**	
$z = 6$		Oldest object seen so far	
$z = 5$		Galaxy clusters form	
$z = 3$	10^6 a	First generation of stars (population II) is formed, starting hydrogen fusion; helium fusion produces carbon, silicon, oxygen	
	$2 \cdot 10^9$ a	First stars explode as supernovae[c]; iron is produced	
$z = 1$	$3 \cdot 10^9$ a	Second generation of stars (population I) appears, and subsequent supernova explosions of the aging stars form the trace elements (Fe, Se, ..) we are made of and blow them into the galaxy	
$4.7 \cdot 10^9$ a		Primitive cloud, made from such explosion remnants, collapses; Sun forms	
$4.6 \cdot 10^9$ a		**Earth** and other planet **formation**; Azoicum starts	
$4.3 \cdot 10^9$ a		Craters form on the planets	
$4.0 \cdot 10^9$ a		Moon forms from material ejected during the collision of a large asteroid with the still liquid Earth	
$4.0 \cdot 10^9$ a		**Archean** eon (Archaeozoicum) starts: bombardment from space stops, Earth's crust solidifies, oldest minerals form, water condenses	
$3.5 \cdot 10^9$ a		Unicellular (microscopic) life appears, stromatolithes	
$2.5 \cdot 10^9$ a		**Proterozoic** eon ('age of first life') starts: atmosphere becomes rich in oxygen due to the activity of microorganisms Ref. 380	
$1 \cdot 10^9$ a		Macroscopic, multicellular life appears	
$800 \cdot 10^6$ a		Earth is completely covered with ice for the first time (reason still unknown) Ref. 381	
600 to $540 \cdot 10^6$ a		Earth is completely covered with ice for the last time	
$540(5) \cdot 10^6$ a		**Paleozoic** era (Palaeozoicum, 'age of old life') starts, after a gigantic ice age: animals appear, oldest fossils (with 540(5) start of Cambrian, 495(5) Ordovician, 440(5) Silurian, 417(5) Devonian, 354(5) Carboniferous and 292(5) Permian period)	

Time from now[a]	Time from big bang[b]	Event	Temperature
$450 \cdot 10^6$ a		Land plants appear	
$370 \cdot 10^6$ a		Wooden trees appear	
$250(5) \cdot 10^6$ a		**Mesozoic** era (Mesozoicum, 'age of middle life', formerly called Secondary) starts: insects and most life forms are exterminated, mammals appear (with 250(5) start of Triassic, 205(4) Jurassic and 142(3) Cretaceous period)	
$150 \cdot 10^6$ a		Continent Pangaea splits into Laurasia and Gondwana	
		The star cluster of the Pleiades forms	
$150 \cdot 10^6$ a		Birds appear	
$142(3) \cdot 10^6$ a		Golden time of dinosaurs (Cretaceous) starts	
$100 \cdot 10^6$ a		Start of formation of Alps, Andes and Rocky mountains	
$65.5 \cdot 10^6$ a		**Cenozoic** era (Caenozoicum, 'age of new life') starts: dinosaurs become extinct due to a comet or asteroid hitting the Earth in the Yucatan, primates appear (with 65.5 start of Tertiary, consisting of Paleogene period with Paleocene, 55.0 Eocene and 33.7 Oligocene epoch, and of Neogene period, with 23.8 Miocene and 5.32 Pliocene epoch; then 1.81 Quaternary period with Pleistocene (or Diluvium) and 0.01 Holocene (or Alluvium) epoch)	
$50 \cdot 10^6$ a		Large mammals appear	
$6-8 \cdot 10^6$ a		Hominids appears	
$5 \cdot 10^6$ a		Homo appears	
500 000 a		Formation of youngest stars in galaxy	
500 000 a		Homo sapiens appears	
100 000 a		Beginning of last ice age	
90 000 a		Homo sapiens sapiens appears	
11 500 a		End of last ice age, start of holocene	
6 000 a		First written texts	
2 500 a		Physics starts	
500 a		Use of coffee, pencil and modern physics starts	
200 a		Electricity use begins	
100 a		Einstein publishes	
10 – 120 a		You were an unicellular being	
present	*c.* $14 \cdot 10^9$ a	You are reading this	$T_\gamma = 2.73$ K, $T_\nu \approx 1.6$ K, $T_b \approx 0$ K
future		You enjoy life; for details and reasons, see page 570	

a. The time coordinate used here is the one given by the coordinate system defined by the microwave background radiation, as explained on page 421. A year is abbreviated a (Latin 'annus').
b. This quantity is not exactly defined since the big bang is not a space-time event. More on the issue on page 954.
c. The history of the atoms shows that we are made from the leftovers of a supernova. We truly are made of *stardust*.

The geological time scale is the one of the International Commission on Stratigraphy; the times are measured through radioactive dating.

Despite its length and its interest, this table has its limitations. For example, what happened elsewhere in the last few thousand million years? There is still a story to be written of which next to nothing is known. For strange reasons, investigations have been rather Earth-centred.

Research in astrophysics is directed at discovering and understanding all phenomena observed in the skies. Here we skip most of this fascinating topic, since as usual, we focus on motion. Interestingly, general relativity allows us to explain many of the general observations about motion in the universe.

The history of space-time

> A number of rabbits run away from a central point in various directions, all with the same speed. While running, one rabbit turns its head, and makes a startling observation. What does it see?

Challenge 808 n

The data showing that the universe is sprinkled with stars all over lead to a simple conclusion: the universe cannot be static. Gravity always changes the distances between bodies; the only exceptions are circular orbits. Gravity also changes the *average* distances between bodies; gravity always tries to collapse clouds. The biggest cloud of all, the one formed by the matter in the universe, must therefore either be collapsing, or still be in expansion.

Ref. 382 The first to dare to draw this conclusion was Aleksander Friedmann.* In 1922 he deduced the detailed evolution of the universe in the case of homogeneous, isotropic mass distribution. His calculation is a classic example of simple but powerful reasoning. For a universe which is homogeneous and isotropic for every point, the line element is given by

Challenge 809 ny

$$ds^2 = c^2 dt^2 - a^2(t)(dx^2 + dy^2 + dz^2) \qquad (340)$$

and matter is described by a density ρ_M and a pressure p_M. Inserting all this into the field

* Aleksander Aleksandrowitsch Friedmann (1888–1925), Russian physicist who predicted the expansion of the universe. Due to his early death from typhus, his work remained almost unknown until Georges A. Lemaître (b. 1894 Charleroi, d. 1966 Leuven), Belgian priest and cosmologist, took it up and expanded it in 1927, focussing, as his job required, on solutions with an initial singularity. Lemaître was one of the propagators of the (erroneous!) idea that the big bang was an 'event' of 'creation' and convinced his whole organization about it. The Friedmann–Lemaître solutions are often erroneously called after two other physicists, who studied them again much later, in 1935 and 1936, namely H.P. Robertson and A.G. Walker.

Page 428, page 429

equations, we get two equations

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\rho_M + \frac{\Lambda}{3} \quad \text{and} \tag{341}$$

$$\ddot{a} = -\frac{4\pi G}{3}\left(\rho_M + 3p_M\right)a + \frac{\Lambda}{3}a \tag{342}$$

which imply

$$\dot{\rho}_M = -3\frac{\dot{a}}{a}\left(\rho_M + p_M\right) . \tag{343}$$

At the present time t_0, the pressure of matter is negligible. (In the following, the index 0 refers to the present time.) In this case, the expression $\rho_M a^3$ is constant in time.

Equations 341 and 342 depend on only two constants of nature: the gravitational constant G, related to the maximum force or power in nature, and the cosmological constant Λ, describing the energy density of the vacuum, or, if one prefers, the smallest force in nature.

Before we discuss the equations, first a few points of vocabulary. It is customary to relate all mass densities to the so-called *critical mass density* ρ_c given by

Challenge 810 ny

$$\rho_c = \frac{3H_0^2}{8\pi G} \approx 8 \pm 2 \cdot 10^{-27}\,\mathrm{kg/m^3} \tag{344}$$

corresponding to about 8, give or take 2, hydrogen atoms per cubic metre. On Earth, one would call this value an extremely good *vacuum*. Such are the differences between everyday life and the universe as a whole. In any case, the critical density characterizes a matter distribution leading to an evolution of the universe just between never-ending expansion and collapse. In fact, this density is the critical one, leading to a so-called *marginal* evolution, only in the case of *vanishing* cosmological constant. Despite this restriction, the term is now used for this expression in all other cases as well. One thus speaks of dimensionless mass densities Ω_M defined as

$$\Omega_M = \rho_0/\rho_c . \tag{345}$$

The cosmological constant can also be related to this critical density by setting

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} = \frac{\Lambda c^2}{8\pi G\rho_c} = \frac{\Lambda c^2}{3H_0^2} . \tag{346}$$

A third dimensionless parameter Ω_K describes the curvature of space. It is defined in terms of the present-day radius of the universe R_0 and the curvature constant $k = (\pm 1, 0)$ as

$$\Omega_K = \frac{-k}{R_0^2\,H_0^2} \tag{347}$$

and its sign is opposite to the one of the curvature k; Ω_K vanishes for vanishing curvature.

Note that a positively curved universe, when homogeneous and isotropic, is necessarily closed and of finite volume. A flat or negatively curved universe with the same matter distribution can be open, i.e. of infinite volume, but does not need to be so. It could be simply or multiply connected. In these cases the topology is not completely fixed by the curvature.

The present time Hubble parameter is defined by $H_0 = \dot{a}_0/a_0$. From equation (341) we then get the central relation

Challenge 811 ny

$$\Omega_M + \Omega_\Lambda + \Omega_K = 1 \,. \quad (348)$$

In the past, when data were lacking, physicists were divided into two camps: the *claustrophobics* believing that $\Omega_K > 0$ and the *agoraphobics* who believe that $\Omega_K < 0$. More details about the measured values of these parameters will be given shortly. The diagram of Figure 199 shows the most interesting ranges of parameters together with the corresponding behaviour of the universe.

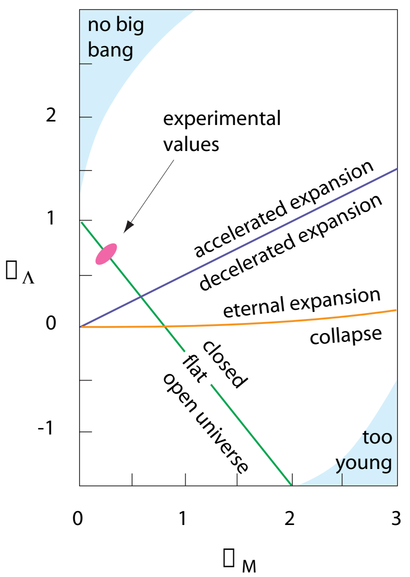


Figure 199 The ranges for the Ω parameters and their consequences

For the Hubble parameter, the most modern measurements give a value of

$$H_0 = 65 \pm 5\,\mathrm{km/sMpc} \approx 2 \cdot 10^{-18}\,/\mathrm{s} \quad (349)$$

which correspond to an age of the universe of 13.5 ± 1.5 thousand million years. In other words, the age deduced from the history of space-time agrees with the age, given above, deduced from the history of stars.

To get a feeling of how the universe evolves, it is customary to use the so-called *deceleration parameter* q_0. It is defined as

$$q_0 = -\frac{\ddot{a}_0}{a_0 H_0^2} = \frac{1}{2}\Omega_M - \Omega_\Lambda \quad (350)$$

The parameter q_0 is positive if the expansion is slowing down, and negative if the expansion is accelerating. These possibilities are also shown in the diagram.

An even clearer way to picture the expansion of the universe for vanishing pressure is to rewrite equation (341) using $\tau = t\,H_0$ and $x(\tau) = a(t)/a(t_0)$, yielding

$$\left(\frac{\mathrm{d}x}{\mathrm{d}\tau}\right)^2 + U(x) = \Omega_K$$
$$\text{with } U(x) = -\Omega_\Lambda x - \Omega_\Lambda x^2 \quad (351)$$

This looks like the evolution equation for the motion of a particle with mass 1, with total energy Ω_K in a potential $U(x)$. The resulting evolutions are easily deduced.

For vanishing Ω_Λ, the universe either expands for ever, or recollapses, depending on

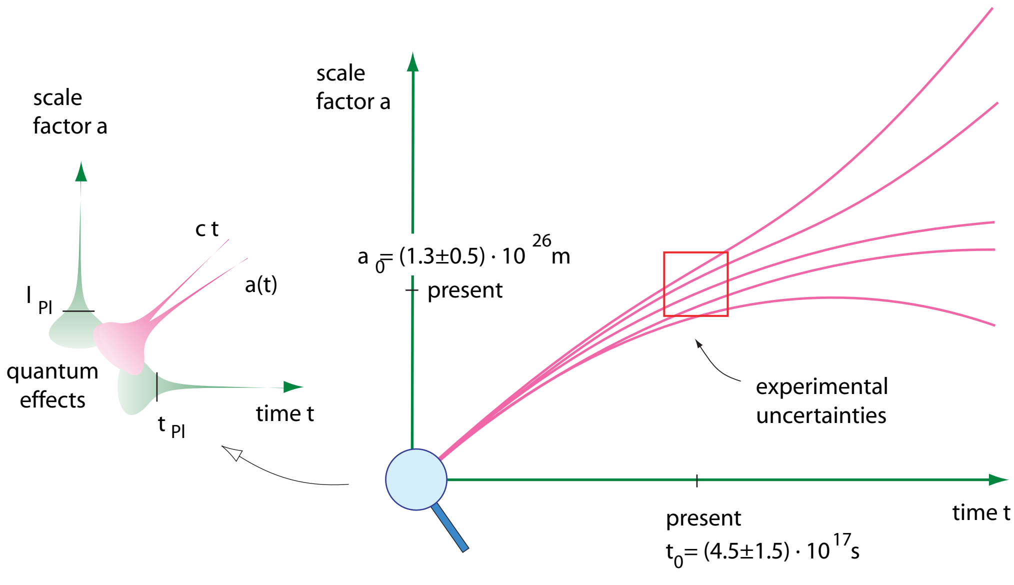


Figure 200 The evolution of the universe's scale R for different values of its mass density

the value of the mass-energy density,

For non-vanishing (positive) Ω_Λ, the potential has exactly one maximum; if the particle has enough energy to get over the maximum, it will accelerate continuously. That is the situation the universe seems to be in today.

For a certain time range, the result is shown in Figure 200. There are two points to be noted: the set of possible curves is described by *two* parameters, not one. In addition, lines cannot be drawn down to the origin of the diagram. There are two main reasons: we do not know the behaviour of matter at very high energy yet, and we do not know the behaviour of space-time at very high energy. We return to this important issue later on.

The main result of Friedmann's work was that a homogeneous and isotropic universe is *not static*: it either expands or contracts. In either case, it has a *finite age*. This profound result took many years to spread around the cosmology community; even Einstein took a long time to get accustomed to it.

Note that due to its isotropic expansion, in the universe there is a preferred reference frame: the frame defined by average matter. The time measured in that frame is the time listed in Table 36 and is the one meant when we talk about the *age* of the universe.

An overview of the possibilities for the long time evolution is given in Figure 201. The evolution can have various outcomes. In the early twentieth century, people decided among them by personal preference. Albert Einstein first preferred the solution $k = 1$ and $\Lambda = a^{-2} = 4\pi G \rho_M$. It is the unstable solution found when $x(\tau)$ remains at the top of the potential $U(x)$.

In 1917, the Dutch physicist Willem de Sitter had found, much to Einstein's personal dismay, that an empty universe with $\rho_M = p_M = 0$ and $k = 1$ is also possible. This type of universe expands for large times. The de Sitter universe shows that in special cases, matter is not needed for space-time to exist.

Challenge 812 ny

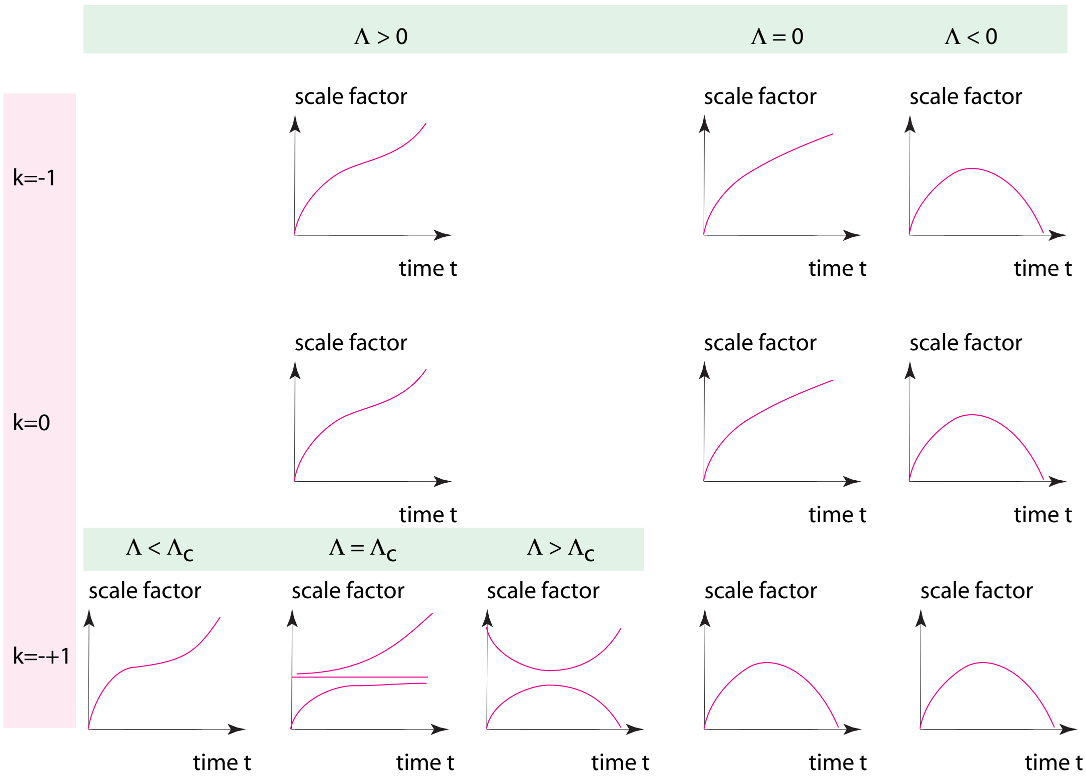


Figure 201 The long-term evolution of the universe's scale factor a for various parameter

Lemaître had found expanding universes for positive mass, and his results were also contested by Einstein in the beginning. When later the first measurements confirmed the calculations, massive and expanding universes became popular. They were promoted to the standard model in textbooks. However, in a sort of collective blindness that lasted from around 1950 to 1990, almost everybody believed that $\Lambda = 0$.* Only towards the end of the twentieth century did experimental progress allow one to make statements free of personal beliefs, as we will find out shortly. But first of all we settle an old issue.

Why is the sky dark at night?

> In der Nacht hat ein Mensch nur ein Nachthemd an,
> und darunter kommt gleich der Charakter.*
> Rober Musil

First of all, the sky is not black at night. It has the same intrinsic colour as during the day, as any long exposure photograph shows. But that colour, like the colour of the sky

Challenge 813 ny

* In this case, for $\Omega_M \geqslant 1$, the age of the universe follows $t_0 \leqslant 2/(3H_0)$, where the limits correspond. For vanishing mass density one has $t_0 = 1/H_0$.

* 'At night, a person is dressed only with a night gown, and under it there is directly the character.' Robert Musil (b. 1880 Klagenfurt, d. 1942 Geneva), German writer.

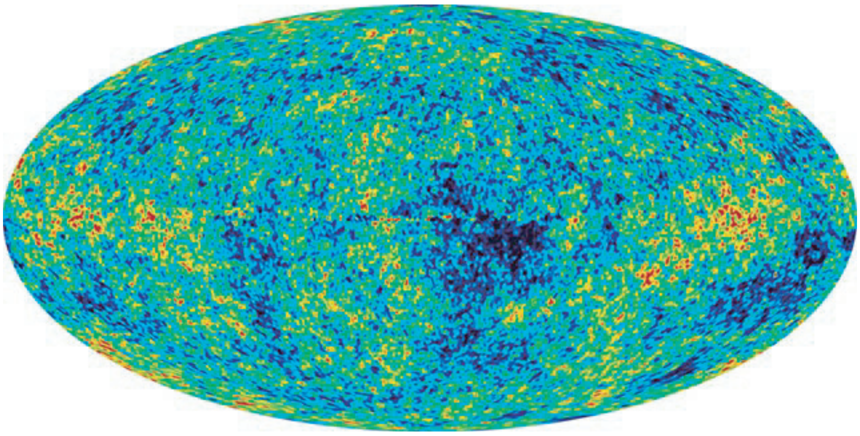


Figure 202 The fluctuations of the cosmic background radiation

during the day, is not due to the temperature of the sky, but to scattered light from the stars. If we look for the real colour of the sky, we need to look for its thermal radiation. Indeed, measurements show that even the empty sky is not completely cold or black at night. It is filled with radiation of around 200 GHz; more precise measurements show that the radiation corresponds to the thermal emission of a body of 2.73 K. This *background radiation* is the thermal radiation left over from the big bang.

Ref. 383 The universe is indeed colder than the stars. But why is this so? If the universe were homogeneous on large scales and infinitely large, it would have an infinite number of stars. Given any direction to look at, we would hit the surface of a star. The night sky would be as bright as the surface of the Sun! Are you able to convince your grandmother about this? Challenge 814 e

In a deep forest, one sees a tree in every direction. Similarly, in a 'deep' universe, we would see a star in every direction. Now, the average star has a surface temperature of about 6000 K. If we would live in a deep and old universe, we would effectively live inside an oven with a temperature of around 6000 K, making it impossible to enjoy ice cream.

This paradox was most clearly formulated in 1823 by the astronomer Wilhelm Olbers.* As he extensively discussed the question, it is also called *Olbers' paradox*. Today we know that even if all matter in the universe were converted into radiation, the universe would still not be as bright as just calculated. In other words, the power and lifetime of stars are way too low to produce the oven brightness just mentioned. So something is wrong. Ref. 384

In fact, two main effects have the power to avoid the contradiction with observations. First, since the universe is finite in age, far away stars are shining for less time. We see them in a younger stage or even during their formation, where they were darker. As a result, the share of brightness of distant stars is smaller than that of nearby stars, so that

* Heinrich Wilhelm Matthäus Olbers (b. 1758 Arbergen, d. 1840 Bremen), astronomer. He discovered two planetoids, Pallas and Vesta, and five comets; he developed the method to calculate parabolic orbits for comets which is still in use today. Olbers also actively supported Friedrich Wilhelm Bessel in his career choice. The paradox is named after Olbers, though others had made similar points before, such as the Swiss astronomer Jean Philippe Loÿs de Cheseaux in 1744 and Johannes Kepler in 1610.

the average temperature of the sky is reduced.*

Second, we could imagine that the radiation of far away stars is shifted to the red and that the volume the radiation must fill is increasing continuously, so that the average temperature of the sky is also reduced. Calculations are necessary to decide which effect is the greater one. This issue has been studied in great detail by Paul Wesson; he explains that the first effect is larger than the second by a factor of about three. We may thus state correctly that the sky is dark at night *mostly* because the universe has a *finite* age. We can add that the sky would be somewhat brighter if the universe were not expanding.

Ref. 385

In addition, the darkness of the sky is possible only because the speed of light is *finite*. Can you confirm this?

Ref. 383 Challenge 816 ny

Finally, the darkness of the sky also tells us that the universe has a *large* (but finite) age. Indeed, the 2.7 K background radiation is that cold, despite having been emitted at 3000 K, because it is red-shifted due to the Doppler effect. Under reasonable assumptions, the temperature T of this radiation changes with the scale factor $R(t)$ of the universe as

Ref. 386

$$T \sim \frac{1}{R(t)} . \tag{352}$$

In a young universe, we would thus not be able to see the stars, even if they existed.

From the brightness of the sky at night, measured to be about $3 \cdot 10^{-13}$ times that of an average star like the Sun, we can deduce something interesting: the density of stars in the universe must be much smaller than in our galaxy. The density of stars in the galaxy can be deduced by counting the stars we see at night. But the average star density in the galaxy would lead to much higher values for the night brightness if it were constant throughout the universe. We can thus deduce that the galaxy is much *smaller* than the universe simply by measuring the brightness of the night sky and by counting the stars in the sky! Can you make the explicit calculation?

Ref. 384 Challenge 817 ny

In summary, the sky is black at night because space-time and matter is of finite, but old age. As a side issue, here is a quiz: is there an Olbers' paradox also for gravitation?

Challenge 818 ny

Is the universe open, closed or marginal?

> - Doesn't the vastness of the universe make you feel small?
> - I can feel small without any help from the universe.
>
> Anonymous

Sometimes the history of the universe is summed up in two words: *bang!...crunch*. But will the universe indeed recollapse or will it expand for ever? Or is it in an intermediate, marginal situation? The parameters deciding its fate are the mass density and cosmological constant.

The main news of the last decade of twentieth-century astrophysics are the experimental results allowing one to determine all these parameters. Several methods are being used. The first method is obvious: determine speed and distance of distant stars. For

* Are you able to explain that the sky is not black because it is painted black or made of black chocolate? Or more generally, that the sky is not made of nor does contain any dark and cold substance, as Olbers himself suggested, and as John Herschel proved wrong in 1848?

Challenge 815 ny

large distances, this is difficult, since the stars get faint. But it has now become possible to search the sky for supernovae, the bright exploding stars, and to determine their distance through their brightness. This is presently being done with the help of computerized searches of the sky, using the largest available telescopes. Ref. 387

A second method is the measurement of the anisotropy of the cosmic microwave background. From the power spectrum as function of the angle the curvature of space-time can be deduced.

A third method is the determination of the mass density using the gravitational lensing effect for the light of distant quasars bent around galaxies or galaxy clusters. Page 432

A fourth method is the determination of the mass density using galaxy clusters. All these measurements are expected to improve greatly in the years to come.

At present, these four completely independent sets of measurements provide the values Ref. 388

$$(\Omega_{\mathrm{M}}, \Omega_{\Lambda}, \Omega_{\mathrm{K}}) \approx (0.3, 0.7, 0.0) \qquad (353)$$

where the errors are of the order of 0.1 or less. The values imply that *the universe is spatially flat, its expansion is accelerating and there will be no big crunch*. However, no definite statement on the topology is possible. We will return to this last issue shortly. Page 434

In particular, the data show that the density of matter, including all dark matter, is only about one third of the critical value.* Twice that amount is given by the cosmological term. For the cosmological constant Λ one gets the value

$$\Lambda = \Omega_{\Lambda} \frac{3H_0^2}{c^2} \approx 10^{-52} \,/\mathrm{m}^2 \,. \qquad (354)$$

This value has important implications for quantum theory, since it corresponds to a vacuum energy density

$$\rho_{\Lambda} c^2 = \frac{\Lambda c^4}{8\pi G} \approx 0.5\,\mathrm{nJ/m^3} \approx \frac{10^{-46}\,(\mathrm{GeV})^4}{(\hbar c)^3} \,. \qquad (355)$$

But the cosmological term also implies a negative vacuum pressure $p_{\Lambda} = -\rho_{\Lambda} c^2$. Inserting this result into the relation for the potential of universal gravity deduced from relativity Page 392

$$\Delta\varphi = 4\pi G(\rho + 3p/c^2) \qquad (356)$$

we get Ref. 389

$$\Delta\varphi = 4\pi G(\rho_{\mathrm{M}} - 2\rho_{\Lambda}) \,. \qquad (357)$$

Challenge 819 ny Thus the gravitational acceleration follows as

* The difference between the total matter density and the separately measurable baryonic matter density, only about one sixth of the former value, is also not explained yet. It might even be that the universe contains matter of a type unknown so far. This issue is called the *dark matter problem*; it is one of the important unsolved questions of cosmology.

$$a = \frac{GM}{r^2} - \frac{\Lambda}{3}c^2 r = \frac{GM}{r^2} - \Omega_\Lambda H_0^2 \, r \, , \tag{358}$$

which shows that a *positive* vacuum energy indeed leads to a *repulsive* gravitational effect. Inserting the mentioned value (354) for the cosmological constant Λ we find that the repulsive effect is small even for the distance between the Earth and the Sun. In fact, the order of magnitude is so much smaller that one cannot hope for a direct experimental confirmation of this deviation from universal gravity at all. Probably astrophysical determinations will remain the only possible ones. A positive gravitational constant manifests itself through a positive component in the expansion rate, as we will see shortly.

Challenge 820 ny

But the situation is puzzling. The origin of this cosmological constant is *not* explained by general relativity. The mystery will be solved only with the help of quantum theory. In any case, the cosmological constant is the first local and quantum aspect of nature detected by astrophysical means.

Why is the universe transparent?

Could the universe be filled with water, which is transparent, as maintained by some popular books in order to explain rain? No. Even if it were filled with air, the total mass would never have allowed the universe to reach the present size; it would have recollapsed much earlier and we would not exist.

Ref. 390

Challenge 821 ny

The universe is thus transparent because it is mostly empty. But *why* is it so empty? First of all, in the times when the size of the universe was small, all antimatter annihilated with the corresponding amount of matter. Only a tiny fraction of matter, which originally was slightly more abundant than antimatter, was left over. This 10^{-9} fraction is the matter we see now. As a consequence, the number of photons in the universe is 10^9 larger than that of electrons or quarks.

Page 872

In addition, 380 000 years after antimatter annihilation, all available nuclei and electrons recombined, forming atoms, and their aggregates, like stars and people. No free charges interacting with photons were lurking around any more, so that from that period onwards light could travel through space like it does today, being affected only when it hits some star or some dust particle.

If we remember that the average density of the universe is $10^{-26}\,\mathrm{kg/m^3}$ and that most of the matter is lumped by gravity in galaxies, we can imagine what an excellent vacuum lies in between. As a result, light can travel along large distances without noticeable hindrance.

But why is the vacuum transparent? That is a deeper question. Vacuum is transparent because it contains no electric charges and no horizons; charges or horizons are indispensable in order to absorb light. Now, quantum theory shows that vacuum does contain so-called *virtual* charges. However, virtual charges have no effects on the transmission of light.

Page 796

The big bang and its consequences

Μελέτη θανάτου. Learn to die.

Plato, *Phaedo*, 81a.

Above all, the big bang model, which is deduced from the colour of the stars and galaxies, states that about fourteen thousand million years ago the whole universe was extremely small. This fact gave the big bang its name. The expression 'big bang' was created (with a Page 642 sarcastic undertone) in 1950 by Fred Hoyle, who by the way never believed that it gives Ref. 391 a correct description of the evolution of the universe. Nevertheless, the term caught on universally. Since the past smallness of the universe cannot itself be checked, we need to look for other, verifiable consequences. The central ones are the following:

- all matter moves away from all other matter;
- there is about 75 mass-% hydrogen and 23 mass-% helium in the universe;
- there is thermal background radiation of about 2.7 K;
- the maximal age for any system in the universe is around fourteen thousand million years;
- there are background neutrinos with a temperature of about 2 K;*
- for non-vanishing cosmological constant, Newtonian gravity is slightly reduced.

All predictions except the last two have been confirmed by observations. Technology probably will not allow us to check them in the foreseeable future; however, there is also no hint of putting them into dispute.

Competing descriptions of the universe have not been successful in matching these Ref. 391 predictions. In addition, theoretical arguments state that with matter distributions such as the observed one, plus some rather weak general assumptions, there is no known way Ref. 392 to avoid a period in the *finite* past in which the universe was extremely small. Therefore it is worth having a close look at the situation.

Was the big bang a big bang?

Was it a kind of explosion? An explosion assumes that some material transforms internal energy into motion of its parts. There has not been any such process in the early history of the universe. In fact, the better description is that space-time is expanding, rather than matter moving. The mechanism and the origin of the expansion is *unknown* at this point of our mountain ascent. Due to the importance of spatial expansion, the whole phenomenon cannot be called an explosion at all. And obviously there neither was nor is any air in interstellar space, so that one cannot speak of a 'bang' in any sense of the term.

Was it big? The visible universe was rather small about fourteen thousand million years ago, much smaller than an atom. In summary, the big bang was neither big nor a bang; but the rest is correct.

Was the big bang an event?

The big bang is a description of what happened in the *whole* of space-time. Despite what is often written in careless newspaper articles, at every moment of the expansion, space

* The theory states that $T_\nu/T_\gamma \approx (4/11)^{1/3}$. These neutrinos appeared about 0.3 s after the big bang.

is always of non-vanishing size; space *never* was a single point. People who pretend this are making ostensibly plausible, but false statements. The big bang is a description of the *expansion* of space-time, not of its beginning. Following the motion of matter back in time, general relativity cannot deduce the existence of an initial singularity. The issue of measurement errors is probably not a hindrance; however, the effect of the nonlinearities in general relativity at situations of high energy densities is not clear.

Most importantly, quantum theory shows that the big bang was *not* a true singularity, as no physical observable, neither density nor temperature, ever reaches an infinitely large
Page 962 (or infinitely small) value. Such values cannot exist in nature.* In any case, there is a general agreement that arguments based on *pure* general relativity alone cannot make correct statements on the big bang. Nevertheless, most newspaper article statements are of this sort.

Was the big bang a beginning?

Asking what was before the big bang is like asking what is north of the North Pole. Since nothing is north of the North Pole, nothing 'was' before the big bang. This analogy could be misinterpreted to imply that the big bang took its start at a single point in time, which of course is incorrect, as just explained. But the analogy is better than it looks; in fact, there is *no* precise North Pole, since quantum theory shows that there is a fundamental indeterminacy as to its position. There is also a corresponding indeterminacy for the big bang.

In fact, it does not take more than three lines to show with quantum theory that time and space are *not* defined either at or near the big bang. We will give this simple argument
Page 923 in the first chapter of the third part of the mountain ascent. The big bang therefore cannot be called a 'beginning' of the universe. There never was a time when the scale factor $R(t)$ of the universe was zero. This conceptual mistake is frequently encountered. In fact, quantum theory shows that near the big bang, events can *neither* be ordered *nor* even be defined. More bluntly, there is *no* beginning; there has never been an initial event or singularity, despite the numerous statements professing the contrary.

Obviously the concept of time is not defined 'outside' or 'before' the existence of the
Ref. 393 universe; this fact was clear to thinkers already over thousand years ago. It is then tempting to conclude that time must have *started*. But as we saw, that is a logical mistake as well: first of all, there is no starting event, and secondly, time does not flow, as clarified
Page 47 already in the beginning of our walk.

A similar mistake lies behind the idea that the universe 'had certain initial conditions.'
Page 139 Initial conditions *by definition* make only sense for objects or fields, i.e. for entities which can be observed from the outside, i.e. for entities which have an environment. The universe does not comply with these requirements; the universe thus cannot have initial conditions. Nevertheless, many people still insist on thinking about the issue; interestingly,
Ref. 394 Stephen Hawking sold millions of books explaining that a description without initial conditions is the most appealing, overlooking that there is no other possibility anyway.**

* Many physicists are still wary to make such strong statements at this point. The first sections of the third
Page 918 part of the mountain ascent give the precise arguments leading to them.

** This statement will still lead to strong reactions among physicists; it will be discussed in more detail in the section on quantum theory.

In summary, the big bang does not contain a beginning nor does it imply one. We will uncover the correct way to think about it in the third part of our mountain ascent.

Does the big bang imply creation?

> [The general theory of relativity produces] universal doubt about god and his creation.
>
> A witch hunter

Page 635

Creation, i.e. the appearance of something out of nothing, needs an existing concept of space and time to make sense. The concept of 'appearance' makes no sense otherwise. But whatever the description of the big bang, be it classical, as in this chapter, or quantum mechanical, as in later ones, this condition is never fulfilled. Even in the present, classical description of the big bang, which gave rise to its name, there is *no* appearance of matter, nor of energy, nor of anything else. And this situation does not change in any later, improved description, as time or space are never defined *before* the appearance of matter.

In fact, all properties of a creation are missing; there is no 'moment' of creation, no appearance from nothing, no possible choice of any 'initial' conditions out of some set of possibilities, and as we will see in more detail later on, not even any choice of particular physical 'laws' from any set of possibilities.

In summary, the big bang does not imply nor harbour a creation process. The big bang was not an event, not a beginning and *not* a case of creation. It is impossible to continue the ascent of Motion Mountain if one cannot accept each of these three conclusions. If one denies them, one has decided to continue in the domain of beliefs, thus effectively giving up on the mountain ascent.

Challenge 822 ny

Note that this requirement is not new. In fact, it was already contained in equation (2) at the start of our walk, as well as in all the following ones. It appears ever more clearly at this point. But what then *is* the big bang? We'll find out in the third part. We now return to the discussion of what the stars can tell us about nature.

Page 50

Why can we see the Sun?

First of all, the Sun is visible because air is transparent. That is not self-evident; in fact air is transparent only to *visible* light and to a few selected other frequencies. Infrared and ultraviolet radiation are mostly absorbed. The reasons lie in the behaviour of the molecules the air consists of, namely mainly nitrogen, oxygen and a few other transparent gases. Several moons and planets in the solar system have opaque atmospheres; we are indeed lucky to be able to see the stars at all.

In fact, even air is not completely transparent; air molecules *scatter* light a little bit. That is why the sky and far away mountains appear blue and sunsets red,* and stars are invisible during daylight. At many frequencies far from visible light the atmosphere is even opaque, as Figure 203 shows.

Secondly, we can see the Sun because the Sun, like all hot bodies, *emits* light. We describe the details of *incandescence*, as this effect is called, below.

Page 564

* Air scattering makes the sky blue also at night, as can be proven by long time exposure cameras; however our eyes are not able to perform this trick, and the low levels of light make it appear black to us.

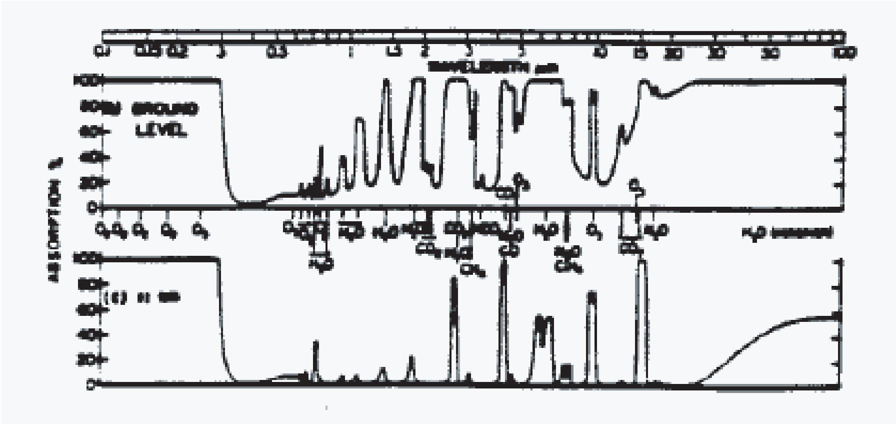


Figure 203 The absorption of the atmosphere

Thirdly, we can see the Sun because we and our environment and the Sun's environment are *colder* than the Sun. In fact, incandescent bodies can be distinguished from their background only if the background is colder. This is a consequence of the properties of incandescent light emission, usually called *black body radiation*. The radiation is material independent, so that for an environment with the same temperature as the body, nothing can be seen at all. Just have a look on the photograph on page 567 as a proof.

Finally, we can see the Sun because it is not a black hole. If it were, it wouldn't emit (almost) any light. Obviously, each of these conditions applies for stars as well. For example, we can only see them because the night sky is black. But then, how to explain the multicoloured sky?

Why are the colours of the stars different?

Stars are visible because they emit visible light. We encountered several important effects which determine colours: the diverse temperatures among the stars, the Doppler shift due to a relative speed with respect to the observer, and the gravitational red-shift.

Not all stars are good approximations of black bodies, so that the black body radiation law sometimes is not an accurate description for their colour. However, most of the stars are reasonable approximations of black bodies. The temperature of a star depends mainly on its size, its mass, its composition and its age, as the astrophysicists are happy to explain. Orion is a good example of a coloured constellation; each star has a different colour. Long term exposure photographs beautifully show this.

Page 565

Ref. 395

Page 68

Note. White dwarfs, or class D stars, are remnants of an imploded star, with a size of only a few tens of kilometres. Not all are white; they can be yellow or red. They form 5% of all stars. None is visible with the naked eye.

The size of all other stars is an independent variable and is sometimes added as roman numeral at the end of the spectral type. (Sirius is an A1V star, Arcturus a K2III star.) Giants and supergiants exist in all classes from O to M.

To accommodate brown dwarfs, two new star classes, L and T, have been proposed.

Table 37 The colour of the stars

Class	Temperature	Example	Position	Colour
O	30 kK	Mintaka	δ Orionis	blue-violet
O	31 ± 10 kK	Alnitak	ζ Orionis	blue-violet
B	22(6) kK	Bellatrix	γ Orionis	blue
B	26 kK	Saiph	κ Orionis	blue-white
B	12 kK	Rigel	β Orionis	blue-white
B	25 kK	Alnilam	ε Orionis	blue-white
B	17(5) kK	Regulus	α Leonis	blue-white
A	9.9 kK	Sirius	α Canis Majoris	blue-white
A	8.6 kK	Megrez	δ Ursae Majoris	white
A	7.6(2) kK	Altair	α Aquilae	yellow-white
F	7.4(7) kK	Canopus	α Carinae	yellow-white
F	6.6 kK	Procyon	α Canis Minoris	yellow-white
G	5.8 kK	Sun	ecliptic	yellow
K	3.5(4) kK	Aldebaran	α Tauri	orange
M	2.8(5) kK	Betelgeuse	α Orionis	red
D	<80 kK	–	–	all

The basic colour determined by temperature is changed by two effects. The first, the
Challenge 823 ny *Doppler red-shift* z, depends on the speed v between source and observer as

$$z = \frac{\Delta\lambda}{\lambda} = \frac{f_S}{f_O} - 1 = \sqrt{\frac{c+v}{c-v}} - 1 . \tag{359}$$

Such shifts play a significant role only for remote, and thus faint, stars visible through the telescope. With the naked eye, Doppler shifts cannot be seen. But Doppler shifts can make distant stars shine in the infrared instead of in the visible domain. Indeed, the highest Doppler shifts observed for luminous objects are larger than 5.0, corresponding to a re-
Challenge 824 ny cessional speed of more than 94 % of the speed of light. Note that in the universe, the red-shift is also related to the scale factor $R(t)$ by

$$z = \frac{R(t_0)}{R(t_{\text{emission}})} - 1 . \tag{360}$$

Light at a red-shift of 5.0 thus was emitted at an age one sixth of the present.

The other colour changing effect, the *gravitational red-shift* z_g, depends on the matter density of the source and is given by

$$z_g = \frac{\Delta\lambda}{\lambda} = \frac{f_S}{f_0} - 1 = \frac{1}{\sqrt{1 - \frac{2GM}{c^2 R}}} - 1 . \tag{361}$$

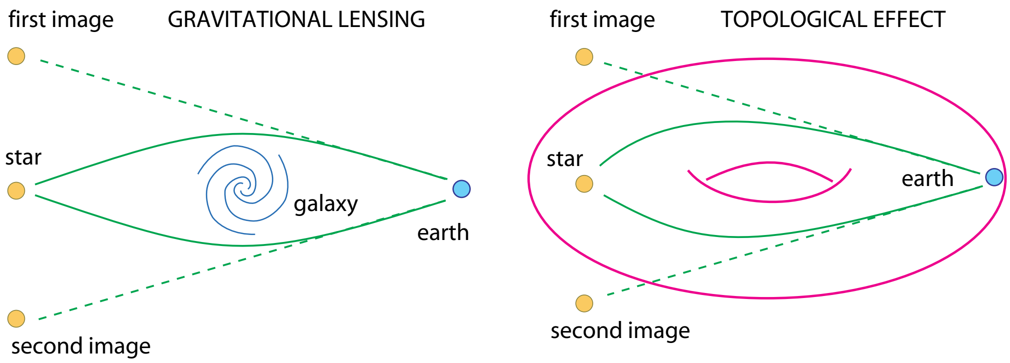


Figure 204 How one star can lead to several images

Challenge 825 ny It is usually quite a bit smaller than the Doppler shift. Can you confirm this?

Other red-shift processes are not known; moreover, such processes would contradict all the properties of nature we know. But the colour issue leads to the next question: Page 440

Are there dark stars?

It could be that some stars are not seen because they are dark. This possibility of dark matter, if widespread, would lead to incorrect matter density estimates for the universe, and thus to incorrect evolution predictions for its fate. This issue is therefore of great interest and hotly debated. It is known that objects more massive than Jupiter but less massive than the Sun can exist in states which do not emit almost any light. They are also called *brown dwarfs*. It is unclear at present how many such objects exist. Many of the so-called extrasolar 'planets' are probably brown dwarfs. The issue is not closed.

Page 441 Another possibility for dark stars are black holes. They are discussed in detail below.

Are all stars different? – Gravitational lenses

Per aspera ad astra.*

Are we sure that at night, two stars are really different? The answer is no. Recently, it was shown that two stars were actually two images of the same object. This was found by comparing the flicker of two different images. It was found that the flicker of one image was exactly the same as the other, just shifted by 423 days. This heroic result was found by the Estonian astrophysicist Jaan Pelt and his research group while observing two quasar images of the system Q0957+561. Ref. 396

The two images are the result of *gravitational lensing*. Indeed, a large galaxy can be seen between the two images, at much smaller distance from the Earth. This effect was already considered by Einstein; however he did not believe that it was observable. The real father of gravitational lensing is Fritz Zwicky, who predicted in 1937 that the effect would be quite common and easy to observe, if lined-up galaxies instead of lined-up stars were considered, as indeed turned out to be the case. Ref. 397

* Through hardship to the stars. A famous Latin motto. Often incorrectly given as 'per ardua at astra'.

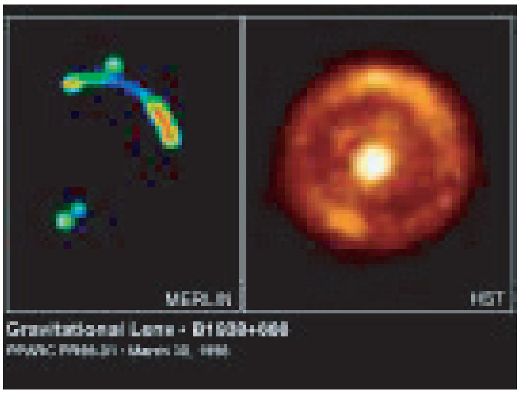


Figure 205 The Zwicky-Einstein ring B1938+666, seen in the radio spectrum (left) and in the optical domain (right)

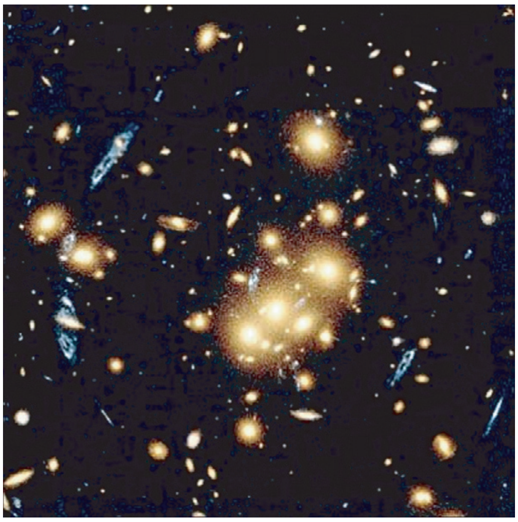


Figure 206 Multiple blue images of a galaxy formed by the yellow cluster CL0024+1654

Interestingly, when the time delay is known, astronomers are able to determine the size of the universe from this observation. Can you imagine how?

Challenge 826 ny

In fact, if the two observed objects are lined up exactly behind each other, the more distant one is seen as *ring* around the nearer one. Such rings have indeed been observed, and the galaxy image around a central foreground galaxy at B1938+666 is one of the most beautiful examples. In 2005, several cases of gravitational lensing by stars have also been observed. More interestingly, three events where one of the two stars has a Earth-mass planet have also been observed. The coming years will surely lead to many additional observations, helped by the sky observation programme in the southern hemisphere that checks the brightness of about 100 million stars every night.

Generally speaking, nearby stars are truly different, but for the distant stars the problem is tricky. For single stars, the issue is not so important, seen overall. Reassuringly, only about 80 multiple star images have been identified so far. But when whole galaxies are seen as several images at once, and several dozens are known so far, we might start to get nervous. In the case of the galaxy cluster CL0024+1654, seven thin, elongated, blue images of the same distant galaxy are seen around the yellow, nearer, elliptical galaxies.

Multiple images can be created not only by gravitational lenses; also the *shape* of the universe could play some tricks.

What is the shape of the universe?

There is a standard explanation to avoid some of the just mentioned problems. The universe in its evolution is similar to the surface of an ever expanding sphere: the surface is finite, but it has no boundary. The universe simply has an additional dimension; therefore its volume is also ever increasing, finite, but without boundary. This statement presupposes that the universe has the same topology, the same 'shape' as that of a sphere with an additional dimension.

Ref. 398 But what is the experimental evidence for this statement? Nothing. Nothing is yet known about the shape of the universe. It is extremely hard to determine it, simply because of its sheer size.

What do experiments say? In the nearby region of the universe, say a few million light years, the topology is simply connected. But for large distances, almost nothing is certain. Maybe research into gamma ray bursts will provide a way to determine topology, as these bursts often originate from the dawn of time, and thus might tell something about the topology.* Maybe even the study of fluctuations of the cosmic background radiation can tell us something. All this research is still in its infancy.

Since little is known, we can ask about the range of possible answers. As just mentioned, in the standard model with $k = 1$, space-time is usually assumed to be a product of linear time, with the topology R of the real line, and a sphere S^3 for space. That is the simplest possible shape, corresponding to a *simply connected* universe. For $k = 0$, the simplest topology of space is three-dimensional real space R^3, and for $k = -1$ it is a hyperbolic manifold H^3.

Page 420 In addition, Figure 199 showed that depending on the value of the cosmological constant, space could be finite and bounded, or infinite and unbounded. In all Friedmann–LeMaître calculations, simple connectedness is usually tacitly assumed, even though it is not at all required.

It could well be that space-time is *multiply* connected, like a higher-dimensional version of a torus. It could also have even more complex topologies.** In these cases, it could even be that the actual number of galaxies is much smaller than the observed number. This situation would correspond to a kaleidoscope, where a few beads produce a large number of images. In addition, topological surprises could also be hidden *behind* the horizon.

In fact, the range of possibilities is not limited to the simply and multiply connected cases suggested by classical physics. An additional and completely unexpected twist

* The story is told from the mathematical point of view by Bob Osserman, *Poetry of the Universe*, 1996.

** The Friedmann–Lemaître metric is also valid for any quotient of the just mentioned simple topologies by a group of isometries, leading to dihedral spaces and lens spaces in the case $k = 1$, to tori in the case $k = 0$, and to *any* hyperbolic manifold in the case $k = -1$. Ref. 399

will appear in the third part of our walk, when quantum theory is included in the investigations. Page 961

What is behind the horizon?

> The universe is a big place; perhaps the biggest.
>
> Kilgore Trout

The horizon is a tricky entity. In fact, all cosmological models show that it moves rapidly away from us. A detailed investigation shows that for a matter dominated universe the horizon moves away from us with a velocity
Ref. 400
Challenge 827 ny

$$v_{\text{horizon}} = 3c \ . \tag{362}$$

A pretty result, isn't it? Obviously, since the horizon does not transport any signal, this is not a contradiction with relativity. But what is behind the horizon?

If the universe were *open* or *marginal*, the matter we see at night is predicted by naively applied general relativity to be a – literally – infinitely small part of all matter existing. Indeed, an open or marginal universe implies that there is an infinite amount of matter behind the horizon. Is such a statement verifiable? In other words, is such a statement a belief or a fact?
Challenge 828 ny

Unfortunately, a *closed* universe fares only slightly better. Matter is still predicted to exist behind the horizon; however, in this case it is only a finite amount.

In short, the standard model of cosmology states that there is a lot of matter behind the horizon. Like most cosmologists, we sweep the issue under the rug and take it up only later in out walk. A precise description of the topic is provided by the hypothesis of inflation.

Why are there stars all over the place? – Inflation

What were the initial conditions of matter? Obviously matter was distributed in a constant density over space expanding with great speed. How could this happen? The person to have explored this question most thoroughly is Alan Guth. So far, we based our studies of the night sky, cosmology, on two observational principles: the isotropy and the homogeneity of the universe. In addition, the universe is (almost) flat. Inflation is an attempt to understand the origin of these observations. Flatness at the present instant of time is strange: the flat state is an unstable solution of the Friedmann equations. Since the universe is still flat after fourteen thousand million years, it must have been even flatter near the big bang.

Ref. 401
Guth argued that the precise flatness, the homogeneity and the isotropy could follow if in the first second of its history, the universe had gone through a short phase of exponential size increase, which he called *inflation*. This exponential size increase, by a factor of about 10^{26}, would homogenize the universe. This extremely short evolution would be driven by a still unknown field, the *inflaton field*. Inflation also seems to describe correctly the growth of inhomogeneities in the cosmic background radiation.

However, so far, inflation poses as many questions as it solves. Twenty years after the initial proposal, Guth himself is sceptical on whether it is a conceptual step forward. The final word on the issue has not been said yet.

Why are there so few stars? – The energy and entropy content of the universe

> Die Energie der Welt ist constant. Die Entropie der Welt strebt einem Maximum zu.*
>
> Rudolph Clausius

The matter-energy density of the universe is near the critical one. Inflation, described in the previous section, is the favourite explanation for this connection. This implies that the actual number of stars is given by the behaviour of matter at extremely high temperatures, and by the energy density left over at lower temperature. The precise connection is still the topic of intense research. But this issue also raises a question about the section quote. Was the creator of the term 'entropy', Rudolph Clausius, right when he made this famous statement? Let us have a look at what general relativity has to say about all this.

In general relativity, a *total* energy can indeed be defined, in contrast to *localized* energy, which cannot. The total energy of all matter and radiation is indeed a constant of motion. It is given by the sum of the baryonic, luminous and neutrino parts:

$$E = E_{\mathrm{b}} + E_{\gamma} + E_{\nu} \approx \frac{c^2 M_0}{T_0} + \ldots + \ldots \approx \frac{c^2}{G} + \ldots . \tag{363}$$

This value is constant only when integrated over the whole universe, not when just the inside of the horizon is taken.**

Many people also add a gravitational energy term. If one tries to do so, one is obliged to define it in such a way that it is exactly the negative of the previous term. This value for the gravitational energy leads to the popular speculation that the *total* energy of the universe might be zero. In other words, the number of stars could be limited also by this relation.

However, the discussion of *entropy* puts a strong question mark behind all these seemingly obvious statements. Many people try to give values for the entropy of the universe. Ref. 402 Some checked whether the relation

$$S = \frac{kc^3}{G\hbar}\frac{A}{4} = \frac{kG}{\hbar c} 4\pi M^2 , \tag{364}$$

Challenge 829 ny which is correct for black holes, also applies to the universe. This assumes that all the matter and all the radiation of the universe can be described by some average temperature. They argue that the entropy of the universe is obviously low, so that there must be some ordering principle behind it. Others even speculate where the entropy of the universe comes from, and whether the horizon is the source for it.

But let us be careful. Clausius assumes, without the slightest doubt, that the universe is a closed system, and thus deduces the above statement. Let us check this assumption. Entropy describes the maximum energy that can be extracted from a hot object. After the

* The energy of the universe is constant. Its entropy tends towards a maximum.
** Except for the case when pressure can be neglected.

discovery of the particle structure of matter, it became clear that entropy is also given by the number of microstates that can make up a specific macrostate. But both definitions make no sense if one applies them to the universe as a whole. There is no way to extract energy from it, and no way to say how many microstates of the universe would look like the macrostate.

The basic reason is the impossibility to apply the concept of *state* to the universe. In the beginning, we defined the state as all those properties of a system which allow one to distinguish it from other systems with the same intrinsic properties, or which differ from one observer to another. You might want to check for yourself that for the universe, such state properties do not exist at all!

Page 38

Challenge 830 ny

We can speak of the state of space-time and we can speak of the state of matter and energy. But we cannot speak of the state of the universe, because the concept makes no sense. If there is no state of the universe, there is no entropy for it. And neither is there an energy value. This is in fact the only correct conclusion one can draw about the issue.

Why is matter lumped?

We are able to see the stars because the universe consists mainly of empty space, in other words, because stars are small and far apart. But why is this the case? Cosmic expansion was deduced and calculated using a homogeneous mass distribution. So why did matter lump together?

It turns out that homogeneous mass distributions are *unstable*. If for any reason the density fluctuates, regions of higher density will attract more matter than regions of lower density. Gravitation will thus have the effect that the denser regions will increase in density, whereas regions of lower density will deplete. Can you confirm the instability, simply by assuming a space filled with dust and $a = GM/r^2$? In summary, even a tiny quantum fluctuation in the mass density will lead, after a certain time, to lumped matter.

Challenge 831 ny

But how did the first inhomogeneities form? That is one of the big problems of modern physics and astrophysics, and there is no accepted answer yet. Several modern experiments are measuring the variations of the cosmic background radiation spectrum with angular position and with polarization; these results, which will be available in the coming years, might provide some information on the way to settle the issue.

Ref. 403

Why are stars so small compared with the universe?

Given that the matter density is around the critical one, the size of stars, which contain most of the matter, is a result of the interaction of the elementary particles composing them. Below we will show that general relativity (alone) cannot explain any size appearing in nature. The discussion of this issue is a theme of quantum theory.

Page 455

Are stars and galaxies moving apart or is the universe expanding?

Can we distinguish between expanding space and galaxies moving apart? Yes, we can. Are you able to find an argument or to devise an experiment to do so?

Challenge 832 ny

The expansion of the universe does not apply to the space on the Earth. The expansion is calculated for a homogeneous and isotropic mass distribution. Matter is not homogeneous nor isotropic inside the galaxy; the approximation of the cosmological principle is not valid down here. It has even been checked experimentally by studying atomic spectra in various places in the solar system that there is *no* Hubble expansion taking place around us. Ref. 404

Is there more than one universe?

'Several' universes might be an option when we study the question whether we see all the stars. But you can check that neither definition of universe given above, be it 'all matter-energy' or 'all matter-energy and all space-time', allows us to answer the question positively. Challenge 833 ny

There is no way to define a plural for universe: either the universe is everything, and then it is unique, or it is not everything, and then it is not the universe. We will discover that quantum theory does not change this conclusion, despite recurring reports to the contrary. Page 757

Why are the stars fixed? – Arms, stars and Mach's principle

The two arms of humans played an important role in discussions about motion, and especially in the development of relativity. Looking at the stars at night, we can make a simple observation, if we keep our arms relaxed. Standing still, our arms hang down. Then we turn rapidly. Our arms lift up. In fact they do so whenever we see the stars turning. Some people have spent a large part of their lives studying this connection. Why?

Ref. 405 Stars and arms prove that motion is obviously relative, not absolute.* This observation leads to two possible formulations of what Einstein called *Mach's principle*.

- *Inertial frames are determined by the rest of the matter in the universe.*

This idea is indeed realized in general relativity. No question about it.

- *Inertia is due to the* interaction *with the rest of the universe.*

This formulation is more controversial. Many interpret this formulation as meaning that the *value of mass* of an object depends on the distribution of mass in the rest of the universe. That would mean that one needs to investigate whether mass is non-isotropic when a large body is nearby. Of course, this question has been studied experimentally; one simply needs to measure whether a particle has the same mass values when accelerated in different directions. Unsurprisingly, to a high degree of precision, no such non-isotropy has been found. Due to this result, many conclude that Mach's principle is wrong. Others conclude with some pain in their stomach that the whole topic is not yet settled. Ref. 406 Ref. 407

But in fact it is easy to see that Mach *cannot* have meant a mass variation at all: one then would also have to conclude that mass should be distance dependent, and that this should be so even in Galilean physics. But this statement is known to be false; nobody in his right mind has ever had any doubts about it. Challenge 834 ny

The whole story is due to a misunderstanding of what is meant by 'inertia': one can interpret it as inertial *mass* or as inertial *motion* (like the moving arms under the stars). There is no evidence that Mach believed either in non-isotropic mass nor in distance-dependent mass; the whole discussion is an example of the frequent plot consisting of

* The original reasoning by Newton and many others around this situation used a bucket and the surface of the water in it; but the arguments are the same.

being proud of not making a mistake which is incorrectly imputed to a supposedly more stupid other person.*

This plot was also applied to Mach's principle. Obviously, inertial effects do depend on the distribution of mass in the rest of the universe. Mach's principle is correct. Mach made some blunders in his life (he is infamous for fighting the idea of atoms until he died, against experimental evidence) but his principle is *not* one of them, in contrast to the story told in many textbooks. But it is to be expected that the myth about the incorrectness of Mach's principle will persist, like that of the derision of Columbus. Ref. 407

In fact, Mach's principle is valuable. As an example, take our galaxy. Experiments show that she is flattened and rotating. The Sun turns around her centre in about 250 million years. Indeed, if the Sun did not turn around the galaxy's centre, we would fall into it in about 20 million years. As the physicist Dennis Sciama pointed out, from the shape of our galaxy we can draw a powerful conclusion: there must be a lot of other matter, i.e. a lot of other stars and galaxies in the universe. Are you able to confirm his reasoning? Challenge 835 ny

At rest in the universe

There is no preferred frame in special relativity, no absolute space. Is the same true in the actual universe? No; there *is* a preferred frame. Indeed, in the standard big-bang cosmology, the average galaxy is at rest. Even though we talk about the big bang, any average galaxy can rightly maintain that it is at rest. Each one is in free fall. An even better realization of this privileged frame of reference is provided by the background radiation.

In other words, the night sky is black because we move with almost no speed through background radiation. If the Earth had a large velocity relative to the background radiation, the sky would be bright even at night. This would be a consequence of Doppler effect for the background radiation. In other words, the fact that the night sky is dark in all directions is a consequence of our slow motion against the background radiation.

The 'slow' motion has the value of about 360 km/s; this value is large in everyday life, but small at cosmic scales. The reason why the galaxy and the solar system move with this 'low' speed across the universe has been already studied in our walk. Can you give a summary? Challenge 836 ny

By the way, is the term 'universe' correct? Does the universe rotate, as its name implies? If by universe one means the whole of experience, the question does not make sense, because rotation is only defined for bodies, i.e. for parts of the universe. However, if by universe one only means 'all matter', the answer *can* be determined by experiments. It turns out that the rotation is extremely small, if there is any: measurements of the cosmic background radiation show that since the universe exists, it cannot have rotated by more than a hundredth of a millionth of a turn! In short, he who talks about the uni'verse' is actually lying. Ref. 408

* For example, at school one usually hears that Columbus was derided because he thought the Earth to be spherical. But he was not derided at all for this reason; there were only disagreements on the *size* of the Earth, and in fact it turned out that his critics were right, and that he was wrong with his own, much too small radius estimate.

Does light attract light?

Another reason that we can see stars is that their light reaches us. But why are travelling light rays not disturbed by gravitation? We know that light is energy and that any energy attracts other energy through gravitation. In particular, light is electromagnetic energy, and experiments showed that all electromagnetic energy is subject to gravitation. Could two light beams that are advancing with a small angle between them focus back together, due to mutual gravitational attraction? That could have measurable and possibly interesting effects on the light observed from distant stars.

The simplest way to explore the issue is to study the following question: Do parallel light beams remain parallel? Interestingly, a precise calculation shows that gravitation Ref. 409 does *not* alter the path of two parallel light beams, even though it *does* alter the path of antiparallel light beams.* The reason is that for parallel beams moving at light speed, the Challenge 837 ny gravitomagnetic component exactly *cancels* the gravitoelectric component.

Since light does not attract light moving along, light is not disturbed by its own gravity during the millions of years that it takes from distant stars to reach us. Light does not attract or disturb light moving nearby. So far, also all known quantum mechanical effects confirm the conclusion.

Does light decay?

In the section on quantum theory we will encounter experiments showing that light is made of particles. It could be that these photons *decay* into some other particle, as yet unknown, or into lower frequency photons. If that actually happened, we would not be able to see distant stars.

Challenge 838 ny But any decay would also mean that light would change its direction (why?) and thus produce blurred images for remote objects. However, no blurring is observed. In addition, the soviet physicist Matvey Bronstein demonstrated in the 1930s that any light decay pro- Ref. 410 cess would have a larger rate for smaller frequencies. Therefore, people checked the shift of radio waves, in particular the famous 21 cm line, and compared it with the shift of light from the same source. No difference was found for all galaxies tested.

People even checked that Sommerfeld's fine-structure constant, the constant of nature Ref. 411 which determines the colour of objects, does not change over time. No sizeable effect could be detected over thousands of millions of years.

Of course, instead of decaying, light could also be *hit* by some hitherto unknown en- Challenge 839 ny tity. But also this case is excluded by the just presented arguments. In addition, these investigations show that there is no additional red-shift mechanism in nature apart from Page 432 Doppler and gravitational red shifts.

The visibility of the stars at night has indeed opened the door to numerous properties of nature. We now continue our mountain ascent with a more general issue, nearer to our quest for the fundamentals of motion.

11. Black holes – falling forever

* Antiparallel beams are parallel beams travelling in opposite directions.

Why study black holes?

> Qui iacet in terra non habet unde cadat.*
> Alanus de Insulis

Black holes are the most extreme case of gravity. Black holes realize nature's limit of length to mass ratios. Black holes produce the highest force value possible in nature; as a result, the realize high space-time curvature values. Therefore, black holes cannot be studied *without* general relativity. In addition, black holes are a central stepping stone towards unification and towards the final description of motion.

Ref. 315 'Black hole' is a short expression for 'gravitationally completely collapsed object'. Strangely enough, for many years their existence was unclear. The available experimental data have led most experts to conclude that there is a black hole at the centre of most galaxies, including our own. Black holes are also suspected at the heart of quasars and of gamma rays bursters. It seems that the evolution of galaxies is strongly tied to the evolution of black holes. In addition, half a dozen smaller black holes have been identified elsewhere in our galaxy. For these and many additional reasons, black holes, the most impressive, the most powerful and the most relativistic systems in nature, are a fascinating subject of study. (Ref. 413, Ref. 412)

Horizons

The escape velocity is the speed needed to launch an projectile in such a way that it never falls back down. The escape velocity depends on the mass and the size of the planet from which the launch takes place. What happens when a planet or star has an escape velocity that is larger than the speed of light c? Such objects were first imagined by the British geologist John Michell in 1784 and independently by the French mathematician Pierre Laplace in 1795, long before general relativity was developed. When they imagined this kind of objects, Michell and Laplace realized something fundamental. Even if an object with such a high escape velocity were a hot bright star, it would appear to be completely *black*. The object would not allow any light to leave it towards the observer; in addition, it would block all light coming form behind it. In 1967, John Wheeler* coined the now standard term *black hole*. (Ref. 414, Ref. 315)

It only takes a short calculation to show that light cannot escape from a mass whenever the radius is smaller than a critical value given by (Challenge 840 ny)

$$R_S = \frac{2GM}{c^2} \tag{365}$$

called the *Schwarzschild radius*. The formula is valid both in universal gravity and in general relativity, provided that in general relativity we take the radius as meaning the circumference divided by 2π. That is exactly the limit value for length to mass ratios in nature.

* 'He who lies on the ground cannot fall down from it.' Alain de Lille, (*c.* 1128–1203).

* John Archibald Wheeler (1911–) US American physicist, important expert on general relativity and author of several excellent textbooks, among them the beautiful John A. Wheeler, *A Journey into Gravity and Space-time*, Scientific American Library & Freeman, 1990, in which he explains general relativity with passion and in detail, but without any mathematics.

For this and other reasons to be given shortly, we will call R_S also the *size* of the black hole of mass M (although properly speaking it is only half the size, and in addition, the term 'size' has to be taken with some care). In principle, an object could be imagined to be smaller than this value; but nobody has yet observed one. In fact we will discover that there is no way to observe an object smaller than the Schwarzschild radius, in the same way that an object moving faster than the speed of light cannot be observed. However, we can observe black holes – the limit case – just as we can observe entities moving at the speed of light.

When a test mass approaches the critical radius R_S^2, two things happen. First, he local proper acceleration for (imaginary) point masses increases without bound. For realistic finite sized objects, the black hole realizes the highest force possible in nature. Something that falls into a black hole cannot be pulled back out. A black hole thus swallows all matter that falls into it without letting anything out. It acts like a cosmic trash can.

At the surface of a black hole, the red-shift factor for a distant observer also increases without bound. The ratio between the two quantities is called the surface gravity of a black hole. It is given by

Challenge 841 ny

$$g_{\text{surf}} = \frac{GM}{R_S^2} = \frac{c^4}{4GM} = \frac{c^2}{2R_S} . \qquad (366)$$

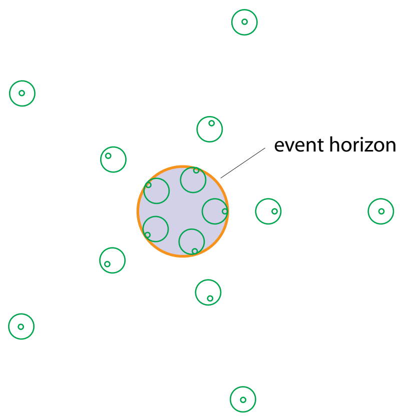


Figure 207 The light cones in the equatorial plane around a non-rotating black hole, seen from above

A black hole thus does not allow any light to leave it.

A surface that realizes the force limit and an infinite red shift makes it is impossible to send light, matter, energy or signals of any kind to the outside world. A black hole is thus surrounded by a horizon. We know that a horizon is a limit surface. A horizon is a limit in two ways. First, a horizon is a limit to communication. Nothing con communicate across the horizon. Second, a horizon is a surface of maximum force and power. These properties are sufficient to answer all questions about the effects of horizons. For example: What happens when a light beam is sent upwards from the horizon? And from slightly above the horizon?

Challenge 842 ny

Black holes, when seen as astronomical objects, are thus different from planets. During the formation of planets, matter clumped together; as soon as it could not be compressed any further, an equilibrium was formed which determined the radius of the planet. That is the same mechanism as when a stone is thrown towards the Earth: it stops falling when it *hits* the floor. A floor is formed whenever matter hits other matter. In the case of a black hole, there is no floor; everything *continues* falling. That is why in Russian, black holes are often called *collapsars*.

The continuous fall takes place when the concentration of matter is so large that it overcomes all those interactions which make matter *impenetrable* in daily life. Already in 1939, Robert Oppenheimer* and Hartland Snyder showed theoretically that a black hole

Ref. 415

* Robert Oppenheimer (1904–1967), important US-American physicist. He can be called the father of

forms whenever a star of sufficient mass stops burning. When a star of sufficient mass stops burning, the interactions that form the 'floor' disappear, and everything continues falling without end.

A *black hole is matter in permanent free fall.* Despite this permanent free fall, its radius for an outside observer remains constant! But that is not all. Due to this permanent free fall, black holes are the only state of matter in thermodynamic equilibrium! Floors and all other states of matter are metastable.

The characterizing property of a black hole is thus its *horizon*. The first time we encountered horizons in special relativity, in the section on accelerated observers. The horizons due to gravitation are similar in all their properties; the section on the maximum force and power has given a first impression. The only difference we have found is due to the neglect of gravitation in special relativity. As a result, horizons in nature cannot be plane, in contrast to the idea suggested by the observations of the imagines point-like observers assumed to exist in special relativity.

Page 306

Both the maximum force principle and the field equations lead to a space-time around a rotationally symmetric, thus non-rotating, and electrically neutral mass described by the Schwarzschild metric

Page 351

$$\mathrm{d}i^2 = (1 - \frac{2GM}{rc^2})\mathrm{d}t^2 - \frac{\mathrm{d}r^2}{1 - \frac{2GM}{rc^2}} - r^2\mathrm{d}\varphi^2/c^2 \ . \tag{367}$$

As mentioned above, r is the circumference divided by 2π and t is the time measured at infinity. No *outside* observer will ever receive any signal emitted from a radius value $r = 2GM/c^2$ or smaller. Indeed, as the proper time i of an observer at radius r is related to the time t of an observer at infinity through

$$\mathrm{d}i = \sqrt{1 - \frac{2GM}{rc^2}}\,\mathrm{d}t \ , \tag{368}$$

we find that an observer at the horizon would have vanishing proper time. In other words, at the horizon the red-shift is infinite. (In fact, the surface of infinite red-shift and the horizon coincide only for non-rotating black holes. For other black holes, such as rotating black holes, the two surfaces are distinct.) Everything happening at the horizon goes on infinitely slowly, when observed by a far away observer. In other words, for a distant observer observing what is going on at the horizon itself, nothing at all ever happens.

In general relativity, horizons of any kind are predicted to be black. Since light cannot escape from them, classical horizons are completely dark. In fact, horizons are the darkest entities imaginable. Nothing in nature is darker than a horizon. Later, we will discover that horizons are not fully black. Nevertheless, we will still be able to state that horizons are the darkest entities in nature.

---

theoretical physics in the USA. He worked on quantum theory and atomic physics. He then headed the development of the nuclear bomb during the second world war. He is also famous for being the most prominent as well as innocent victim of one of the greatest witch-hunts that were organized in his home country. See also the http://www.nap.edu/readingroom/books/biomems/joppenheimer.html web site.

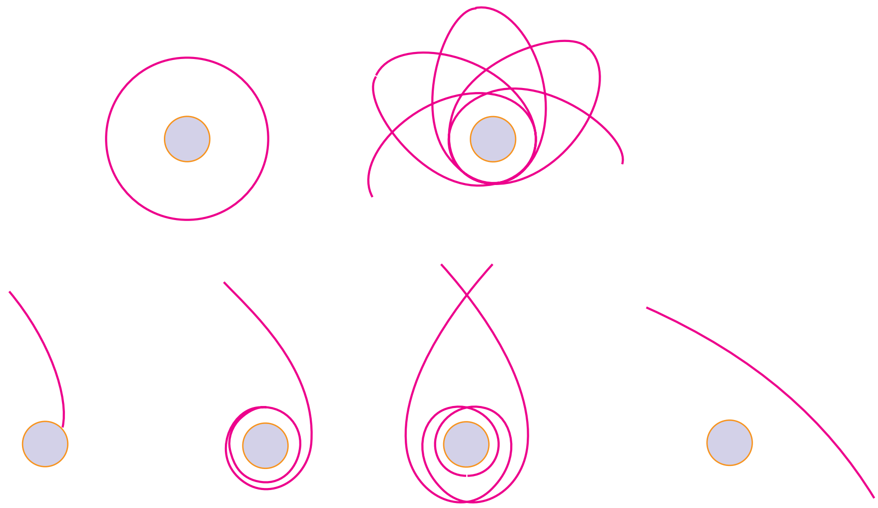


Figure 208 Motion of uncharged objects around a non-rotating black hole

Orbits

Ref. 410 Since black holes curve space-time strongly, a body moving near a black hole behaves in more complicated ways than in the case of universal gravity. In universal gravity, paths are either ellipses, parabolas, or hyperbolas; all these are plane curves. It turns out that paths lie in a plane only near *non-rotating* black holes.*

Around non-rotating black holes, also called *Schwarzschild black holes*, circular paths are impossible for radii less than $3R_S/2$ (can you show why?) and are unstable to perturbations from there up to a radius $3R_S$. Only at larger radii circular orbits are stable. Around black holes, there are no elliptic paths; the corresponding rosetta path is shown in Figure 208. Such a path shows the famous periastron shift in all its glory. (Challenge 844 ny)

Note that the potential around a black hole is not appreciably different from $1/r$ for distances above about fifteen Schwarzschild radii. For a black hole of the mass of the Sun, that would be 42 km from its centre; at the distance of the Earth, we would not be able to note any difference for the path of the Earth around the Sun. (Challenge 845 ny)

Several times in our adventure we mentioned that gravitation is characterized by its tidal effects. Black holes show extreme properties also in this aspect. If a cloud of dust falls into a black hole, the size of the cloud increases when falling into it, until the cloud envelops the whole horizon. In fact, the result is valid for any extended body. This property of black holes will be of importance later on, when we will discuss the size of elementary particles.

For falling bodies coming from infinity, the situation near black holes is even more interesting. Of course there are no hyperbolic paths, only trajectories similar to hyper-

* For such paths, Kepler's rule connecting the average distance and the time of orbit

$$\frac{GMt^3}{(2\pi)^2} = r^3 \qquad (369)$$

still holds, provided the proper time and the radius measured by a far away observer is used. (Challenge 843 ny)

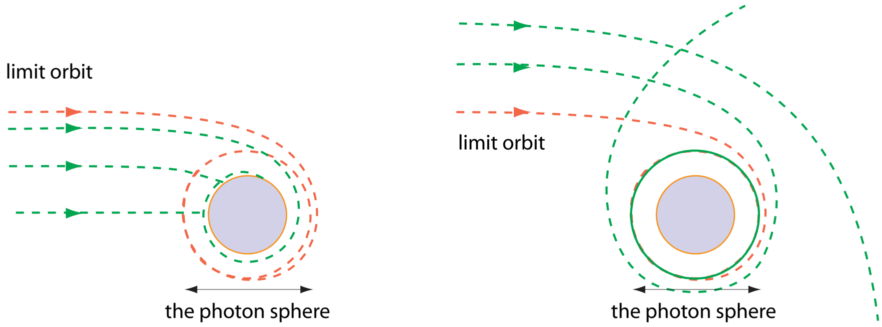


Figure 209 Motion of *light* passing near a non-rotating black hole

bolas for bodies passing far enough. But for small, but not too small impact parameters, a body will make a number of turns around the black hole, before leaving again. The number of turns increases beyond all bounds with decreasing impact parameter, until a value is reached at which the body is captured into an orbit at a radius of $2R$, as shown in Figure 208. In other words, this orbit *captures* incoming bodies if they reach it below a certain critical angle. For comparison, remember that in universal gravity, no capture exists. At still smaller impact parameters, the black hole swallows the incoming mass. In both cases, capture and deflection, a body can make several turns around the black hole, whereas in universal gravity, it is impossible to make more than *half* a turn around a body.

The most absurd looking orbits though are those (purely academic) orbits correspond- ing to the parabolic case of universal gravity. In summary, relativity changes the motions due to gravity quite drastically. Challenge 846 ny

Around *rotating* black holes, the orbits of point masses are even more complex than those shown in Figure 208; for bound motion for example, the ellipses do not stay in one plane, but also change – due to the Lense–Thirring effect – the plane in which they lie, leading to extremely involved orbits in three dimensions filling the space around the black hole.

For *light* passing a black hole, the paths are equally interesting, as shown in Figure 209. There are no qualitative differences with the case of rapid particles, as relativity requires. For a non-rotating black hole, the path obviously lies in a single plane. Of course, if light passes sufficiently nearby there is strong bending of light, as well as capture. Again, light can also make one or several turns around the black hole before leaving or being captured. The limit between the two cases is the path in which light moves in a circle around a black hole, at $3R/2$. If we would be located on that orbit, we would see the back of our head by looking forward! However, this orbit is unstable. The space containing all orbits inside the circular one is called the *photon sphere*. The photon sphere thus divides paths leading to capture from those leading to infinity. As a note, there is no *stable* orbit for light around a black hole at all. Are there any rosetta paths for light around a black hole? Challenge 847 ny Challenge 848 ny

For light around a *rotating* black hole, paths are much more complex. Already in the equatorial plane there are two possible circular light paths, namely a smaller one in direction of the rotation, and a larger one in the opposite direction. Challenge 849 ny

For *charged* black holes, the orbits for falling charged particles are even more complex.

The electrical field lines need to be taken into account; several fascinating effects appear with no correspondence in usual electromagnetism, such as effects similar to electrical versions of the Meissner effect. The whole field is still partly unexplored and is one of today's research themes in general relativity.

Hair and entropy

How is a black hole characterized? It turns out that black holes have no choice for their size, their shape, their colour, their magnetic field and all their material properties to be discussed later on. They all follow from the few properties characterizing them, namely their mass M, their angular momentum J, and their electrical charge Q.* All other properties are uniquely determined by them.** It is as though one could deduce every characteristic of a woman only by her size, her waist and her height, following Wheeler's colourful language. Physicists also say that black holes 'have no hair,' meaning that (classical) black holes have no other degrees of freedom. This expression also was introduced by Ref. 418 Wheeler.*** This was shown by Israel, Carter, Robinson and Mazur; they showed that for a black hole with given mass, angular momentum and charges, there is only *one* possible black hole. Ref. 419 (However, the uniqueness theorem is not valid any more if the black hole carries nuclear quantum numbers, such as weak or strong charges.)

In other words, independently of how the black hole has formed, independently of which material and composition was used when building it, the final result does not depend on those details. Black holes all have identical composition, or better, they have no composition at all (at least classically).

The mass of a black hole is not restricted by general relativity. It may be as small as that of a microscopic particle and as large as many million solar masses. But for their angular momentum J and for their electric charge Q the situation is different. A rotating black hole has a maximum possible angular momentum and a maximum possible electrical (and magnetic) charge.**** The limit in angular momentum appears as its perimeter may Challenge 850 ny not move faster than light. Also the electrical charge is limited. The two limits are not independent; they are related by

$$\left(\frac{J}{cM}\right)^2 + \frac{GQ^2}{4\pi\varepsilon_0 c^4} \leqslant \left(\frac{GM}{c^2}\right)^2 . \quad (370)$$

* There are other entities encountered so far with the same reduced number of characteristics: particles. More on the connection between black holes and particles will be uncovered in the third part of our mountain ascent.

** Mainly for marketing reasons, non-rotating and electrically neutral black holes are often called *Schwarzschild* black holes; Ref. 416 uncharged and rotating ones are often called *Kerr* black holes, after Roy Kerr, who discovered the corresponding solution of Einstein's field equations in 1963. Electrically charged, but non-rotating black holes are often called *Reissner–Nordström black holes*, after the German physicist Hans Reissner and the Finnish physicist Gunnar Nordström. The general case, charged and rotating, is sometimes Ref. 417 named after Kerr and Newman.

Ref. 315 *** Wheeler writes that he was inspired by the difficulty to distinguish bald men; however, Feynman, Ruffini and others had a clear anatomical image when they stated that 'black holes, in contrast to their surroundings, have no hair.'

**** More about the still hypothetical magnetic charge later on. It enters like an additional type of charge into all expressions in which electric charge appears.

Challenge 851 ny

The limit simply follows from the limit on length to mass ratios at the basis of general relativity. Rotating black holes realizing the limit (370) are called *extremal* black holes. The limit (370) also implies that the horizon radius of a general black hole is given by

$$r_{\mathrm{h}} = \frac{GM}{c^2}\left(1+\sqrt{1-\frac{J^2c^2}{M^4G^2}-\frac{Q^2}{4\pi\varepsilon_0 GM^2}}\right) \qquad (371)$$

For example, for a black hole with the mass and half the angular momentum of the Sun, namely $2\cdot10^{30}$ kg and $0.45\cdot10^{42}\,\mathrm{kg\,m^2/s}$, the charge limit is about $1.4\cdot10^{20}$ C.

Ref. 420

How does one distinguish rotating from non-rotating black holes? First of all by the *shape*. Non-rotating black holes must be spherical (any non-sphericity is radiated away as gravitational waves) and rotating black holes have a slightly flattened shape, uniquely determined by their angular momentum. Due to their rotation, their surface of infinite gravity or infinite red-shift, called the *static limit*, is different from their (outer) horizon. The region in between is called the *ergosphere*; as a misnomer, it is *not* a sphere. (It is called this way because, as we will see shortly, it can be used to extract energy from the black hole.) The motion of bodies between the ergosphere and the horizon can be quite complex. It suffices to mention that rotating black holes drag any infalling body into an orbit around them, in contrast to non-rotating black holes, which swallow them. In other words, rotating black holes are not really 'holes' at all, but rather black vortices.

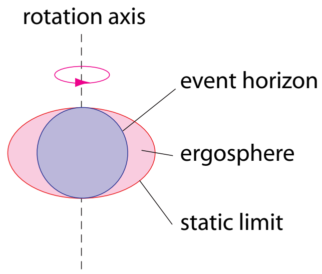


Figure 210 The ergosphere of a rotating black hole

Challenge 852 ny

The distinction between rotating and non-rotating black holes also appears in the horizon surface area. The (horizon) surface A of a non-rotating and uncharged black hole is obviously related to its mass M by

$$A = \frac{16\pi G^2}{c^4}M^2\,. \qquad (372)$$

The surface-mass relation for a rotating and charged black hole is more complex; it is given by

$$A = \frac{8\pi G^2}{c^4}M^2\left(1+\sqrt{1-\frac{J^2c^2}{M^4G^2}-\frac{Q^2}{4\pi\varepsilon_0 GM^2}}\right) \qquad (373)$$

where J is the angular momentum and Q the charge. In fact, the relation

$$A = \frac{8\pi G}{c^2}Mr_{\mathrm{h}} \qquad (374)$$

is valid for *all* black holes, even if charged and rotating. Obviously, in the case of electrically charged black holes, the rotation also produces a magnetic field around them. This is in contrast with non-rotating black holes which cannot have a magnetic field.

Black holes as energy sources

Can one extract energy from a black hole? Roger Penrose discovered that this is possible for *rotating* black holes. A rocket orbiting a rotating black hole in its ergosphere could switch its engines on and then would get hurled into outer space at tremendous velocity, much greater than what the engines could have produced by themselves. In fact, the same effect is used by rockets on the Earth as well and is the reason that all satellites orbit the Earth in the same direction; it would cost much more fuel to let them turn the other way.* Anyway, the energy gained by the rocket is lost by the black hole, which thus slows down and would lose some mass; on the other hand, the mass increases due to the exhaust gases falling into the black hole. This increase always is larger or at best equal to the loss due to rotation slowdown. The best one can do is to turn the engines on exactly at the horizon; then the horizon area of the black hole stays constant, and only its rotation is slowed down.**

Ref. 421

As a result, for a neutral black hole *rotating* with its maximum possible angular momentum, $1 - 1/\sqrt{2} = 29.3\,\%$ of its total energy can be extracted through the Penrose process. For black holes rotating more slowly, the percentage is obviously smaller.

Challenge 854 ny

For *charged* black holes, such irreversible energy extraction processes are also possible. Can you think of a way? Using expression (370), we find that up to 50 % of the mass of a non-rotating black hole can be due to its charge. In fact, in the second part of the mountain ascent we will encounter a process which nature seems to use quite frequently.

Challenge 855 ny
Challenge 856 ny
Page 815

The Penrose process allows one to determine how angular momentum and charge increase the mass of a black hole. The result is the famous mass-energy relation

Ref. 422

$$M^2 = \frac{E^2}{c^4} = (m_{\mathrm{irr}} + \frac{Q^2}{16\pi\varepsilon_0 G m_{\mathrm{irr}}})^2 + \frac{J^2}{4m_{\mathrm{irr}}^2}\frac{c^2}{G^2} = (m_{\mathrm{irr}} + \frac{Q^2}{8\pi\varepsilon_0 \rho_{\mathrm{irr}}})^2 + \frac{J^2}{\rho_{\mathrm{irr}}^2}\frac{1}{c^2} \qquad (375)$$

which shows how the electrostatic and the rotational energy enter the mass of a black hole. In the expression, m_{irr} is the *irreducible mass* defined as

$$m_{\mathrm{irr}}^2 = \frac{A(M,0,0)}{16\pi}\frac{c^4}{G^2} = \left(\rho_{\mathrm{irr}}\frac{c^2}{2G}\right)^2 \qquad (376)$$

and ρ_{irr} is the *irreducible radius*.

These investigations showed that there is no process which *decreases* the horizon area and thus the irreducible mass or radius of the black hole. People have checked this in all ways possible and imaginable. For example, when two black holes merge, the total area increases. One calls processes which keep the area and energy of the black hole constant *reversible*, and all others irreversible. In fact, the area of black holes behaves like the *entropy* of a closed system: it never decreases. That the area in fact *is* an entropy was first

* And it would be much more dangerous, since any small object would hit such an against-the-stream satellite with about 15.8 km/s, thus transforming any small object into a dangerous projectile. In fact, any power wanting to destroy satellites of the enemy would simply have to load a satellite with nuts or bolts, send it into space the wrong way and distribute the bolts into a cloud. It would make satellites impossible for many decades to come.

Challenge 853 ny

** It is also possible to extract energy from rotational black holes through gravitational radiation.

Ref. 423 stated in 1970 by Jakob Bekenstein. He deduced that only when an entropy is ascribed to a black hole, was it possible to understand where the entropy of all the material falling into it was collected.

The black hole entropy is only a function of the mass, the angular momentum and the charge of a black hole. You might want to confirm Bekenstein's deduction that the entropy Challenge 857 ny S is proportional to the horizon area. Later it was found, using quantum theory, that

$$S = \frac{A}{4}\frac{kc^3}{\hbar G} = \frac{A\,k}{4\,l_{\mathrm{Pl}}^2} . \qquad (377)$$

This famous relation needs quantum theory for its deduction, as the absolute value of entropy, as for any other observable, is never fixed by classical physics alone. We will Page 817 discuss the entropy expression later on in our mountain ascent.

If black holes have an entropy, they also must have a temperature. If they have a temperature, they must shine. Black holes thus cannot be black! The last conclusion was proven by Stephen Hawking in 1974 with extremely involved calculations. However, it could have been deduced already in the 1930s, with a simple Gedanken experiment that Page 813 we will present later on. You might want to think about the issue, asking and investigating what strange consequences would appear if black holes had no entropy. Black hole radiation is a further, though tiny (quantum) mechanism for energy extraction, and is applic- Page 813 able even for non-rotating, uncharged black holes. The interesting connections between black holes, thermodynamics, and quantum theory will be presented in the second part Challenge 858 ny of our mountain ascent. Can you imagine other mechanisms that make black hole shine?

Paradoxes, curiosities and challenges

> Tiens, les trous noirs. C'est troublant.*
>
> Anonyme

Black holes show many counter-intuitive results. We first have a look at the classical ef- Page 820 fects. The quantum effects are left for later on.

- Following universal gravity, a black hole would allow that light climbs upwards from its surface and then falls back down. In general relativity, a black hole does not allow light Challenge 859 ny to climb up at all; it can only fall. Can you confirm this?
- What happens to a person falling into a black hole? An outside observer gives a clear answer: the falling person *never* arrives there since she needs an infinite time to reach the Challenge 860 ny horizon. Can you confirm this result? The falling observer however, reaches the horizon Challenge 861 ny in a *finite* amount of his own time. Can you calculate it?

This result is surprising, as it means that for an outside observer in a universe with *finite* age, black holes cannot have formed yet! At best, we can only observe systems busy forming black holes. In a sense, it might be correct to say that black holes do not exist. However, black holes could have existed right from the start in the fabric of space-time. On the other hand, we will find out later why this is impossible. In other words, it is important to keep in mind that the idea of black hole is a limit concept.

* No translation possible. Traduttore, traditore.

Independently of this last issue, we can confirm that in nature, the length to mass ratio always follows

$$\frac{L}{M} \geqslant \frac{4G}{c^2} . \tag{378}$$

▪ Interestingly, the *size* of a person falling into a black hole is experienced in vastly different ways by the falling person and the one staying outside. If the black hole is large, the infalling observer feels almost nothing, as the tidal effects are small. The outside observer makes a startling observation: he sees the falling person *spread* all over the horizon of the black hole. *Infalling, extended bodies cover the whole horizon.* Can you explain the result, e.g. by using the limit on length to mass ratios?

Challenge 862 ny

This strange result will be of importance later on in our exploration, and lead to important results for the size of point particles.

▪ An observer near a (non-rotating) black hole, or in fact near any object smaller than 7/4 times its gravitational radius, can even see the complete *back* side of the object, as shown in Figure 211. Can you imagine how the image looks? Note that in addition to the paths shown in Figure 211, light can also turn several times around the black hole before hitting its surface! Therefore, such an observer sees an infinite number of images of the black hole. The formula for the angular size of the innermost image was given above.

Challenge 863 ny

Page 361

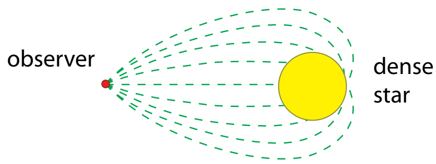


Figure 211 Motion of some light rays from a dense body to an observer

In fact, gravity has the effect to allow the observation of more than half a sphere of *any* object. In everyday life the effect is not so large; for example, light bending allows us to see about 50.0002 % of the surface of the Sun.

▪ A mass point inside the smallest circular path of light around a black hole, at $3R/2$, cannot stay in a circle, because in that region, something strange happens. A body which circles another in everyday life always feels a tendency to be pushed outwards; this centrifugal effect is due to the inertia of the body. But at values below $3R/2$, a circulating body is pushed *inwards* by its inertia. There are several ways to explain this paradoxical effect. The simplest is to note that near a black hole, the weight increases faster than the centrifugal force, as you may want to check yourself. Only a rocket with engines switched on and pushing towards the sky can orbit a black hole at $3R/2$.

Ref. 424

Challenge 864 ny

▪ By the way, how can gravity or an electrical field come out of a black hole, if no signal and no energy can leave it?

Challenge 865 n

▪ Do *white holes* exist, i.e. time inverted black holes, in which everything flows out of instead of into some bounded region?

Challenge 866 ny

Challenge 867 ny

▪ Show that a cosmological constant Λ leads to the following metric for a black hole:

$$\mathrm{d}\tau^2 = \frac{\mathrm{d}s^2}{c^2} = \left(1 - \frac{2GM}{rc^2} - \frac{\Lambda}{3}r^2\right)\mathrm{d}t^2 - \frac{\mathrm{d}r^2}{c^2 - \frac{2GM}{r} - \frac{\Lambda c^2}{3}r^2} - \frac{r^2}{c^2}\mathrm{d}\varphi^2 . \tag{379}$$

▪ In quantum theory, the *gyromagnetic ratio* is an important quantity for any rotating charged system. What is the gyromagnetic ratio for rotating black holes?

Challenge 868 ny

▪ A large black hole is, as the name implies, black. Still, it can be seen. If we would travel towards it in a space ship, we would note that the black hole is surrounded by a bright circle, like a thin halo. The ring at the radial distance of the photon sphere is due to those photons which come from other luminous objects, then circle the hole and finally, after one or several turns, end up in our eye. Can you confirm this result? Challenge 869 n

Challenge 870 ny
▪ Do moving black holes Lorentz-contract? Black holes do shine a little bit; it is true that the images they form are complex, as the light can turn around them a few times, before reaching the observer. In addition, the observer has to be far away, so that curvature has small effects. All these effects can be taken into account; nevertheless, the question remains subtle. The reason is that the concept of Lorentz contraction makes no sense in general relativity, as the comparison with the uncontracted situation is difficult to define precisely.

Challenge 871 ny
▪ Can you confirm that black holes provide a limit to power? Power is energy change over time. General relativity limits power to $P = c^5/4G$. In other words, no engine in nature can provide more than $0.92 \cdot 10^{52}$ W or $1.2 \cdot 10^{49}$ horsepower.*

Formation of and search for black holes

How might black holes form? At present, at least three mechanisms are distinguished; the question is still a hot subject of research. First of all, black holes could have formed during the early stages of the universe. These *primordial black holes* might grow through *accretion*, i.e. through the swallowing of nearby matter and radiation, or disappear through one of the mechanisms to be studied later on. Ref. 425 Page 815

Of the *observed* black holes, the so-called *supermassive* black holes are found at the centre of every galaxy studied so far. They have masses in the range from 10^6 to 10^9 solar masses and contain about 0.5 % of the mass of a galaxy. They are conjectured to exist at the centre of all galaxies and seem to be related to the formation of galaxies themselves. Supermassive black holes are supposed to have formed through the collapse of large dust collections, and to have grown through subsequent accretion of matter. The latest ideas imply that these black holes accrete a lot of matter in their early stage; the matter falling in emits lots of radiation and thus would explain the brightness of quasars. Later on, the accretion calms down, and the less spectacular Seyfert galaxies form. It may even be that the supermassive black hole at the centre of the galaxy triggers the formation of stars. Still later, these supermassive black holes almost get dormant, like the one in the centre of our own galaxy.

Ref. 426
On the other hand, black holes can form when old massive stars *collapse*. It is estimated that when stars with at least three solar masses burn out their fuel, part of the matter will collapse into a black hole. Such *stellar* black holes have a mass between one and a hundred solar masses; they can also continue growing through subsequent accretion. This situation provided the first candidate ever, Cygnus X-1, which was discovered in 1971.

Recent measurements suggest also the existence of *intermediate* black holes, with masses around thousand solar masses or more; their formation mechanisms and formation conditions are still unknown.

The search for black holes is a popular sport among astrophysicists. The conceptually simplest way to search for them is to look for strong gravitational fields. But only double

* This statement is not yet found in the literature on general relativity. So beware.

stars allow one to measure fields directly, and the strongest ever measured gravitational field so far is 30 % of the theoretical maximum value. Another way is to look for strong gravitational lenses, and try to get a mass to size ratio pointing to a black hole. Still another way is to look at the dynamics of stars near the centre of galaxies. Measuring their motion, one can deduce the mass of the body they orbit. The most favourite method to search for black holes is to look for extremely intense X-ray emission from point sources, using space-based satellites or balloon based detectors. If the distance to the object is known, its absolute brightness can be deduced; if it is above a certain limit, it must be a black hole, since normal matter cannot produce an unlimited amount of light. The method is being perfected with the aim to achieve the direct observation of energy disappearing into a horizon. This might have been observed recently.

Ref. 427
Ref. 428

To sum up the experimental situation, measurements show that in all galaxies studied so far – more than a dozen – a supermassive black hole seems to be located at their centre. The masses vary; the black hole at the centre of our own galaxy has about 2.6 million solar masses. The central black hole of the galaxy M87 has 3 thousand million solar masses.

Ref. 413

About a dozen stellar black holes between 4 and 20 solar masses are known in the rest of our own galaxy, all discovered in the years after 1971, when Cygnus X-1 was found. In the year 2000, intermediate mass black holes have also been found. Astronomers are also studying how large numbers of black holes in star clusters behave, how often they collide and what sort of measurable gravitational waves these collisions produce. The list of discoveries and the related results are expected to expand dramatically in the coming years.

Ref. 413

Singularities

Solving the equations of general relativity for various initial conditions, one finds that a cloud of dust usually collapses to a *singularity*, i.e. to a point of infinite density. The same conclusion appears when one follows the evolution of the universe backwards in time. In fact, Roger Penrose and Stephen Hawking have proven several mathematical theorems on the necessity of singularities for many classical matter distributions. These theorems assume only the continuity of space-time and a few rather weak conditions on the matter in it. The theorems state that in expanding systems such as the universe itself, or in collapsing systems such as black holes in formation, events with infinite matter density should exist somewhere in the past, respectively in the future. This result is usually summarized by saying that there is a mathematical proof that the universe started in a singularity.

Ref. 429

In fact, the derivation of the initial singularities assumes a hidden, but strong property of matter: that dust particles have no proper size. In other words, it is assumed that that dust particles are singularities themselves. Only with this assumption can one deduce the existence of initial singularities. However, we have seen that the maximum force principle can be reformulated as a minimum size principle for matter. The argument that leads to an initial singularity of the universe is thus not correct; it is based on an incorrect assumption. The experimental situation is clear: there is overwhelming evidence for an early state of the universe that was extremely hot and dense; but there is *no* evidence for infinite temperature or density.

Mathematically influenced researchers distinguish two types of singularities: with and without a horizon. The latter ones, the so-called *naked* singularities, are especially strange; for example, a tooth brush can fall into a naked singularity and disappear without leaving any trace. Since the field equations are time invariant, we could thus expect that every now and then, naked singularities emit tooth brushes. (Can you explain why dressed singularities are less dangerous?) Challenge 872 ny

To avoid the spontaneous appearance of tooth brushes, over the years many people have tried to discover some theoretical principles forbidding the existence of naked singularities. It turns out that there are two such principles. The first is the maximum force or maximum power principle we encountered above. The maximum force implies that no infinite force values appear in nature; in other words, there are no naked singularities in nature. This statement is often called *cosmic censorship*, Obviously, if general relativity would not be the correct description of nature, naked singularities could still appear. Cosmic censorship is thus still discussed in research articles. The experimental search for naked singularities has not yielded any success; in fact, there is not even a candidate observation for a dressed singularity. But also the case for 'dressed' singularities is weak. Since there is no way to interact with anything behind a horizon, it is futile to discuss what happens there. There is no way to prove that behind a horizon a singularity exists. Dressed singularities are entities of religion, not of physics. Ref. 430

In fact, there is another principle preventing singularities, namely *quantum theory*. Whenever we encounter a prediction of an infinite value, we have extended our description of nature to a domain for which it was not conceived. To speak about singularities, one must assume the applicability of pure general relativity to very small distances and very high energies. As will become clear in the next two parts of the book, nature does not allow this; the combination of general relativity and quantum theory shows that it makes no sense to talk about 'singularities' nor about what happens 'inside' a black hole horizon. The reason is that time and space are not continuous at smallest distances. * Page 962 Page 923

A quiz: is the universe a black hole?

Could it be that we live inside a black hole? Both the universe and black holes have horizons. Even more interesting, the horizon distance r_0 of the universe is about

$$r_0 \approx 3ct_0 \approx 4 \cdot 10^{26}\,\mathrm{m} \tag{380}$$

and its matter content is about

$$m_0 \approx \frac{4\pi}{3}\rho_o r_0^3 \quad \text{whence} \quad \frac{2Gm_0}{c^2} = 72\pi G\rho_0 c t_0^3 = 6 \cdot 10^{26}\,\mathrm{m} \tag{381}$$

for a density of $3 \cdot 10^{-27}\,\mathrm{kg/m^3}$. Thus we have

$$r_0 \approx \frac{2Gm_0}{c^2} \tag{382}$$

* Many physicists are still wary to make such strong statements at this point, especially – of course – all those who claim that space and time are continuous even down to the smallest distances. The part on quantum theory and the first sections of the third part of the mountain ascent give the precise arguments leading to the opposite conclusion. Page 918

Challenge 873 ny similar to the black hole relation $r_S = 2Gm/c^2$. Is this a coincidence? No, it is not; all systems with high curvature more or less obey the relation. But are we nevertheless falling
Challenge 874 ny into a large black hole? You can find out by yourself.

12. Does space differ from time?

Tempori parce.

Seneca*

People in a bad mood say that time is our master. Nobody says that of space. Time and space are obviously different in everyday life. But what is the precise difference between them in general relativity? And do we need them at all? In general relativity it is assumed that we live in a (pseudo-Riemannian) space-time of variable curvature. The curvature is an observable and is related to the distribution and motion of matter and energy in the way described by the field equations.

However, there is a fundamental problem. The equations of general relativity are invariant under numerous transformations which *mix* the coordinates x_0, x_1, x_2 and x_3. For example, the viewpoint transformation

$$\begin{aligned} x_0' &= x_0 + x_1 \\ x_1' &= -x_0 + x_1 \\ x_2' &= x_2 \\ x_3' &= x_3 \end{aligned} \qquad (383)$$

is allowed in general relativity, and leaves the field equations invariant. You might want
Challenge 875 ny to search for other examples.

The consequence is clearly in sharp contrast with everyday life: diffeomorphism invariance makes it *impossible* to distinguish space from time *inside* general relativity. More explicitly, the coordinate x_0 cannot simply be identified with the physical time t, as implicitly done up to now. This identification is only possible in *special* relativity. In special relativity the invariance under Lorentz (or Poincaré) transformations of space and time singles out energy, linear and angular momentum as the fundamental observables. In general relativity, there is *no* metric isometry group; consequently, there are *no* basic physical observables singled out by their characteristic of being conserved. But invariant quantities are necessary for communication! In fact, we can *talk* to each other only because we live in an approximately *flat* space-time. If the angles of a triangle would not add up to 180 degrees, we could not communicate, since there would be no invariant quantities.

How did we sweep this problem under the rug so far? We used several ways. The simplest was to always require that in some part of the situation under consideration space-time is our usual flat Minkowski space-time, where x_0 can be set equal to t. This requirement can be realized either at infinity, as we did around spherical masses, or in zeroth approximation, as we did for gravitational radiation and for all other perturbation calculations. In this way, the free mixing of coordinates is eliminated and the otherwise missing invariant quantities appear as expected. This pragmatic approach is the usual way

* 'Care about time.' Lucius Annaeus Seneca (*c.* 4 BCE–65), *Epistolae* 88, 39.

out of the problem. In fact, there are otherwise excellent texts on general relativity that preclude any deeper questioning of the issue.

Ref. 386

A common variation of this trick is to let the distinction 'sneak' into the calculations by the introduction of matter and its properties, or by the introduction of radiation. Both matter and radiation distinguish between space and time simply by their presence. The material properties of matter, for example their thermodynamic state equations, always distinguish space and time. Radiation does the same, by its propagation. Obviously this is true also for those special combinations of matter and radiation called clocks and metre bars. In fact, the method of introducing matter is the same as the one introducing Minkowski space-time, if one looks closely: matter properties are always defined using space-time descriptions of special relativity.*

Still another variation of the pragmatic approach is the use of the cosmological time coordinate. An isotropic and homogeneous universe does have a preferred time coordinate, namely the one used in all the tables on the past and the future of the universe. Also this method is in fact a combination of the previous two.

But we are on a special quest here. We want to *understand* motion, not only to calculate its details. We want a *fundamental* answer, not a pragmatic one. And for this we need to know how the positions x_i and time t are connected, and how we can define invariant quantities. The question also prepares us for the moment when gravity is combined with quantum theory, as we will do in the third part of our mountain ascent.

A fundamental solution requires a description of clocks together with the system under consideration, and a deduction of how the reading t of the clock relates to the behaviour of the system in space-time. But we know that any description of a system requires measurements, e.g. in order to determine the initial conditions. And initial conditions require space and time. We enter a vicious circle; that is precisely what we wanted to avoid in the first place.

A suspicion arises. Does a fundamental difference between space and time exist at all? Let us have a tour of the various ways to investigate the question.

Can space and time be measured?

In order to distinguish space and time in general relativity, we must be able to measure them. But already in the section on universal gravity we had mentioned the impossibility of measuring lengths, times and masses with gravitational effects alone. Does this situation change in general relativity? Lengths and times are connected by the speed of light, and in addition lengths and masses are connected by the gravitational constant. Despite this additional connection, it takes only a moment to convince oneself that the problem persists. In fact, we need *electrodynamics* to solve it. Only using the electromagnetic charge e can we form length scales, of which the simplest one is given by

Page 241

Ref. 432

$$l_{\mathrm{scale}} = \frac{e}{\sqrt{4\pi\varepsilon_0}} \frac{\sqrt{G}}{c^2} \approx 1.4 \cdot 10^{-36}\,\mathrm{m} \,. \qquad (384)$$

* We note something astonishing here: the inclusion of some condition at small distances (matter) has the same effect as the inclusion of some condition at infinity. Is this a coincidence? We will come back to this issue in the third part of the mountain ascent.

Challenge 876 ny
Page 971

Page 486 Here, e is the elementary charge and ε_0 the permittivity of free space. In fact, only *quantum mechanics* provides a real solution to this issue, as can be seen by rewriting the elementary charge e as the combination of nature's fundamental constants using

$$e = \sqrt{4\pi\varepsilon_0 c\hbar\alpha} \,. \tag{385}$$

Here, $\alpha \approx 1/137.06$ is the fine-structure constant that characterizes the strength of electromagnetism. This connection changes expression (384) into

$$l_{\text{scale}} = \sqrt{\frac{\alpha\hbar G}{c^3}} = \sqrt{\alpha}\, l_{\text{Pl}} \,. \tag{386}$$

The expression shows that every length measurement is based on the electromagnetic coupling constant α and on the Planck length. Of course, the same is valid for time and mass measurements as well. There is thus no way to define or measure lengths, times and masses in general relativity alone.* Therefore, the answer to the section title is negative. The next question then is whether in general relativity space and time are really necessary or are merely convenient constructs.

Challenge 877 e

Are space and time necessary?

Ref. 433 Robert Geroch answers this question in a beautiful five-page article. He explains how to formulate the general theory of relativity without the use of space and time, by taking as starting point the physical observables only.

He starts with the set $\{a\}$ of all observables. Among them there is one, called v, which stands out. It is the only observable which allows one to say that for any two observables a_1, a_2 there is a third one a_3, for which

$$(a_3 - v) = (a_1 - v) + (a_2 - v) \,. \tag{387}$$

Such an observable is called the *vacuum*. Once such an observable is known, Geroch shows how to use it to construct the derivatives of observables. Then the so-called Einstein algebra can be built, which comprises the whole of general relativity.

Usually one describes motion by deducing space-time from matter observables, by calculating the evolution of space-time, and then by deducing the motion of matter following from it. Geroch's description shows that the middle step, the use of space and time, is not necessary.

Indirectly, the principle of maximum force makes the same statement. General relativity can be derived from the existence of limit values for force or power. Space and time are only needed to translate this principle into consequences for real-life observers. Space and time are tools for this translation.

Ref. 431 * In the past, John Wheeler used to state that his *geometrodynamic clock*, a device which measures time by bouncing light back and forward between two parallel mirrors, was a counterexample; that is not correct, however. Can you confirm this?

Challenge 878 n

We conclude that it is possible to formulate general relativity without the use of space and time. Since both are unnecessary, it is unlikely that there is a fundamental difference between them. Nevertheless, one difference between time and space is well-known.

Do closed timelike curves exist?

Is it possible that the time coordinate behaves, at least in some regions, like a torus? Is it possible, like in space, to come back in time to where we have started? The question has been studied in great detail. The standard reference is the text by Hawking and Ellis; they list the various properties of space-time which are mutually compatible or exclusive. Among others, they find that space-times which are smooth, globally hyperbolic, oriented and time-oriented do not contain any such curves. It is usually assumed that these properties apply to the observed universe, so that nobody expects to observe closed timelike curves. Indeed, no candidate has ever been named.

Ref. 392

Later on, we will show that also searches at the microscopic scale have led to negative results on this question.

Page 823

The impossibility of closed timelike curves seems to point to a difference between space and time. But in fact, this difference is only apparent. All these investigations are based on the behaviour of matter. Thus these arguments imply a specific answer right from the start and do not allow one to search for it. In short, also this topic cannot help to decide whether space and time differ. Let us look at the issue in another way.

Is general relativity local? – The hole argument

When Albert Einstein developed general relativity, he had quite some trouble with diffeomorphism invariance. Most startling is his famous *hole argument*, better called the *hole paradox*. Take the situation shown in Figure 212, in which a mass deforms the space-time around it. Einstein imagined a small region of the vacuum, the *hole*, which is shown as a small ellipse. What happens if we somehow change the curvature inside the hole while leaving the situation outside it unchanged, as shown in the inset of the picture?

Ref. 434

On one hand, the new situation is obviously physically different from the original one, as the curvature inside the hole is different. This difference thus implies that the curvature outside a region does not determine the curvature inside it. That is extremely unsatisfactory. Worse, if we generalize this operation to the time domain, we get the biggest nightmare possible in physics: determinism is lost.

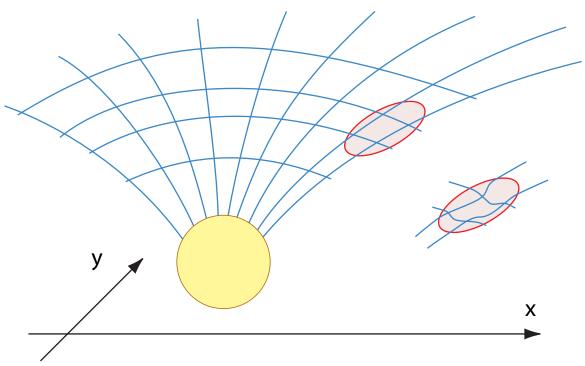


Figure 212 A 'hole' in space

On the other hand, general relativity is diffeomorphism invariant. The deformation shown in the figure is a diffeomorphism. The situation must be physically equivalent to the original situation.

Who is right? Einstein first favoured the first point of view, and therefore dropped the whole idea of diffeomorphism invariance for about a year. Only later did he understand that the second assessment is correct, and that the first statement makes a fundamental mistake. Indeed, the first opinion arrives at the conclusion that the two situations are physically different because it assumes an independent existence of the coordinate axes x and y, as shown in the figure. But during that deformation, the coordinates x and y automatically change as well, so that there is *no* physical difference between the two situations.

The moral of the story is that *there is no difference between space-time and gravitational field*. Space-time is a quality of the field, as Einstein put it, and not an entity with separate existence, as assumed in the graph. Coordinates have no physical meaning; only distances (intervals) in space and time have one. In particular, diffeomorphism invariance proves that *there is no flow of time.* Time, like space, is only a relational entity: time and space are relative; they are not absolute.

The relativity of space and time also has practical consequences. For example, it turns out that many problems in general relativity are equivalent to the Schwarzschild situation, even though they appear completely different at first sight. As a result, researchers have 'discovered' the Schwarzschild solution (of course with different coordinate systems) over twenty times, often thinking that they had found a new, unknown solution. The topic has a startling consequence.

Is the Earth hollow?

The *hollow Earth hypothesis*, i.e. the conjecture that we live on the *inside* of a sphere, was popular in paranormal circles around the year 1900, and still remains so among certain
Page 55
eccentrics today, especially in Britain, Germany and the US. They maintain that the solid Earth *encloses* the sky, together with the Moon, the Sun and the stars. Most of us are fooled by education into the usual description, because we are brought up to believe that light travels in straight lines. Get rid of this belief, it is said, and the hollow Earth appears in all its glory.

Interestingly, the reasoning is correct. There is *no way* to disprove this sort of description of the universe. In fact, as the great Austrian physicist Roman Sexl used to
Ref. 435
explain, the diffeomorphism invariance of general relativity even proclaims the equivalence between the two views. The fun starts when either of the two camps wants to tell the other that *only* its own description is correct. You might check that any such argument is wrong; it is fun to slip into the shoes of such an eccentric and to defend the hollow
Challenge 879 e
Earth hypothesis against your friends. Explaining the appearance of day and night, of the horizon and of the satellite images of the Earth is easily done. Explaining what happened during the flight to the Moon is also droll. You can drive many bad physicists crazy in this way. The usual description and the hollow Earth description are exactly equivalent. Are you able to confirm that even quantum theory, with its introduction of length scales
Challenge 880 n
into nature, does not change the situation?

All these investigations show that diffeomorphism invariance is not an easy symmetry to swallow. But it is better to get used to it now, as the rest of our adventure will bring even more surprises. Indeed, in the third part of our walk we will discover that there is an even larger symmetry of nature that is similar to the change in viewpoint from the

Figure 213 A model of the hollow Earth theory

hollow Earth view to the standard view. This symmetry, space-time duality, is not only valid for distances measured from the centre of the Earth, but for distances measured from any point in nature. Just be patient. Page 971

Are space, time and mass independent?

We conclude from this short discussion that there does not seem to be a fundamental distinction between space and time in general relativity. Pragmatic distinctions, using matter, radiation or space-time at infinity are the only possible ones.

In the third part of our adventure we will discover that even the inclusion of quantum theory is consistent with this view. We will show explicitly that no distinction is possible in principle. We will discover that mass and space-time are on an equal footing and that in a sense, particles and vacuum are made of the same substance. All distinctions between space and time turn out to be possible only at low, daily life energies. Page 935

Page 144 In the beginning of our mountain ascent we found that we needed matter to define space and time. Now we even found that we need matter to *distinguish* space and time. Similarly, in the beginning we found that space and time are required to define matter; now we found that we even need *flat* space-time to define it. Page 398

In summary, general relativity does not solve the circular reasoning we discovered in Galilean physics. General relativity even makes the issue less clear than before. Continuing the mountain ascent is really worth the effort. To increase our understanding, we now go to the limit case of gravitation.

13. General relativity in ten points – a summary for the layman

Sapientia felicitas.*

General relativity is the final description of *paths of motion*, or if one prefers, of *macroscopic motion*. General relativity describes how the observations of motion of *any* two observers are related to each other, and also describes motion due to gravity. In fact, general relativity is based on the following observations:

- All observers agree that there is a 'perfect' velocity in nature, namely a common maximum energy velocity relative to matter. The preferred velocity is realized by massless radiation, such as light or radio signals.
- All observers agree that there is a 'perfect' force in nature, a common maximum force that can be realized or measured by realistic observers. The perfect force is realized on event horizons.

These two statements contain the full theory of relativity. From these two central facts we deduce:

- Space-time consists of events in 3+1 *continuous dimensions*, with a curvature varying from point to point. The curvature can be deduced from distance measurements among events or from tidal effects. We thus live in a pseudo-riemannian space-time. Measured times, lengths and curvatures vary from observer to observer.
- Space-time and space is *curved near mass and energy*. The average curvature at a point is determined by the energy–momentum density at that point and described by the field equations. When matter and energy move, the space curvature moves along with them. A built-in delay in this renders faster than light transport of energy impossible. The proportionality constant between energy and curvature is so small that the curvature is not observed in everyday life, but only its indirect manifestation, namely gravity.
- Space is also *elastic*; it prefers being flat. Being elastic, it can wiggle also independently of matter; one then speaks of gravitational radiation or of gravity waves.
- Freely falling matter moves along *geodesics*, i.e. along paths of maximal length in curved space-time; in space this means that light bends when it passes near large masses by twice the amount predicted by universal gravity.
- To describe gravitation one *needs* curved space-time, i.e. general relativity, *at the latest* whenever distances are of the order of the Schwarzschild radius $r_S = 2Gm/c^2$. When distances are large than this value, the relativistic description with gravity and gravitomagnetism (frame-dragging) is sufficient. When distances are even larger, the description by universal gravity, namely $a = Gm/r^2$, together with flat Minkowski space-time, will do as approximation.
- Space and time are not distinguished globally, but only locally. *Matter* is required to perform the distinction.

In addition, all matter and energy we observe in the sky provide two observations:

- On cosmological scale, everything moves away from everything else: the universe is *expanding*. This expansion of space-time is described by the field equations.
- The universe has a *finite age*; the finiteness is the reason for the darkness at night. A horizon limits the measurable space-time intervals to about fourteen thousand million years.

* 'Wisdom is happiness.' This is also the motto of Oxford university.

The accuracy of the description

Was general relativity worth the effort? The discussion of its accuracy is most conveniently split into two sets of experiments. The first set is given by measurements of how *matter moves*. Do objects really follow geodesics? As summarized in Table 38, all experiments agree with theory within measurement errors, i.e. at least within 1 part in 10^{12}. In short, the way matter falls is indeed described by general relativity in all details.
Ref. 436
Ref. 437

The second set of measurements checks the dynamics of space-time itself. Does *space-time move* following the field equations of general relativity? In other words, is space-time really bent by matter in the way the theory predicts? Many experiments have been performed, near and far from Earth, both in weak and in strong fields. All agree with the predictions within errors. However, the best measurements so far have only about 3 significant digits. Note that even though numerous experiments have been performed, there are only few *types* of tests, as Table 38 shows; in the past, the discovery of a new type has always meant fame and riches. Most sought after, of course, is the direct detection of gravitational waves.
Ref. 436, Ref. 437
Challenge 881 ny

Another comment on Table 38 is in order. After many decades in which all measured effects were only of order v^2/c^2, several so-called *strong field effects* in pulsars allowed us to reach order v^4/c^4. Soon a few effects of this order should also be detected even inside the solar system, using high-precision satellite experiments. The present crown of all measurements, the gravity wave emission delay, is the only v^5/c^5 effect measured so far.
Ref. 436
Page 373

The difficulty to achieve high precision for space-time curvature measurements is the reason that mass is measured with balances, always (indirectly) using the prototype kilogram in Paris, instead of defining some standard curvature and fixing the value of G. Indeed, no terrestrial curvature experiment has ever been carried out. Also in this domain a breakthrough would make the news. At present, any terrestrial curvature method would not even allow one to define a kilogram of gold or of oranges with a precision of a single kilogram!

Another possible check of general relativity is the search for alternative descriptions of gravitation. Quite a number of competing theories of gravity have been formulated and studied, but none is in agreement with all experiments.
Ref. 437, Ref. 439

In summary, as Thibault Damour likes to explain, general relativity is at least 99.999 999 999 9 % correct concerning the motion of matter and energy, and at least 99.9 % correct about the way matter and energy curve and move space-time. No exceptions, no anti-gravity and no unclear experimental data are known. All motion on Earth and in the skies is described by general relativity. The importance of the achievement of Albert Einstein cannot be understated. We note that general relativity has not been tested for microscopic motion. In this context, *microscopic motion* is any example of motion for which the action is around the quantum of action, namely 10^{-34} Js. This issue is central to the third and last part of our adventure.
Ref. 436

Research in general relativity and cosmology

Despite all these successes, research in general relativity is more intense than ever.*
Ref. 441

Table 38 Present types of tests of general relativity

Measured effect	Confirmation	Type	Reference
Equivalence principle	10^{-12}	motion of matter	Ref. 329, Ref. 436
$1/r^2$ dependence (dimensionality of space-time)	10^{-10}	motion of matter	Ref. 438
Time independence of G	10^{-19} /s	motion of matter	Ref. 436
Red-shift (light & microwaves on Sun, Earth, Sirius)	10^{-4}	space-time curvature	Ref. 307, Ref. 306, Ref. 436
Perihelion shift (four planets, Icarus, pulsars)	10^{-3}	space-time curvature	Ref. 436
Light deflection (light, radio waves around Sun, stars, galaxies)	10^{-3}	space-time curvature	Ref. 436
Time delay (radio signals near Sun, near pulsars)	10^{-3}	space-time curvature	Ref. 436
Gravitomagnetism (Earth, pulsar)	10^{-1}	space-time curvature	Ref. 331
Geodesic effect (Moon, pulsars)	10^{-1}	space-time curvature	Ref. 349, Ref. 436
Gravity wave emission delay (pulsars)	10^{-3}	space-time curvature	Ref. 436

- The description of collisions and of many body problems, around the motion of stars, neutron stars and black holes, with its richness of behaviour, helps astrophysicists to improve their understanding of what they observe in their telescopes. Ref. 442
- The study of the early universe and of elementary particle properties, with topics such as *inflation*, a short period of accelerated expansion during the first few seconds, is still an important topic of investigation. Ref. 443
- The study of chaos in the field equations is of fundamental interest in the study of the early universe, and may be related to the problem of galaxy formation, one of the biggest open problems in physics. Ref. 444
- Gathering data about galaxy formation is the main aim of many satellite systems and purpose-build telescopes. The main focus is the search for localized cosmic microwave background anisotropies due to protogalaxies. Ref. 445
- The determination of the cosmological parameters, such as the matter density, the curvature and the vacuum density, is a central effort of modern astrophysics. Ref. 388
- Astrophysicists regularly discover new phenomena in the skies. For example, the various types of gamma ray bursts, X-ray bursts and optical bursts are still not completely understood. Gamma ray bursts, for example, can be as bright as 10^{17} usual stars combined; however, they last only a few seconds. More details on this research is given later on. Ref. 446 Page 815
- A computer database of all solutions of the field equations is being built. Among others, researchers are checking whether they really are all different from each other. Ref. 447

* There is even a free and excellent internet based research journal, called *Living Reviews in Relativity*, to be found at the http://www.livingreviews.org website.

▪ The inclusion of torsion into field equations, a possible extension of the theory, is one of the promising attempts to include particle spin into general relativity. Ref. 448

▪ Studying solutions with non-trivial topology, such as wormholes and particle-like solutions, is a fascinating field of enquiry, also related to string theory. Ref. 449

▪ Other formulations of general relativity, describing space-time with quantities other than the metric, are continuously developed, in the hope to clarify the relation to the quantum world. In fact, the unification of quantum physics and general relativity, the topic of the third part of this mountain ascent, will occupy researchers for many years to come. Ref. 450

▪ Finally, the teaching of general relativity, which for many decades has been hidden behind Greek indices, differential forms and other antididactic methods, will benefit greatly from future improvements focusing more on the physics and less on the formalism. Ref. 451

In short, general relativity is still an extremely interesting field of research and important discoveries are still expected.

Could general relativity be different?

The constant of gravitation provides a limit for the density and the acceleration of objects, as well as for the power of engines. We based all our deductions on its invariance. Is it possible at all that the constant of gravitation G changes from place to place or that it changes with time? The question is tricky. On first sight, the answer is a loud 'Yes, of course! Just experience what happens when the value of G is changed in formulae.' However, this statement is wrong, as it was wrong for the speed of light c. Page 311

Since the constant of gravitation enters our definition of gravity and acceleration, and thus enters, even if we do not notice it, the construction of all rulers, all measurement standards and all measuring set-ups, there is *no way* to detect whether its value actually varies. No imaginable experiment could detect a variation. Every measurement of force is, whether we like it or not, a comparison with the limit force. There is no way, in principle, to check the invariance of a standard. This is even more astonishing because measurements of this types are regularly reported, and this chapter is no exception. But the result of any such experiment is easy to predict: no change will ever be found. Challenge 882 ny Page 462

Could the number of space dimension be different from 3? This issue is quite involved. For example, three is the smallest number of dimensions for which a vanishing Ricci tensor is compatible with non-vanishing curvature. On the other hand, more than three dimensions give deviations from the inverse square 'law' of gravitation. So far, there are no data pointing in this direction.

Could the equations of general relativity be different? Despite their excellent fit with experiment, there is one issue that still troubles some people. The rotation speed of matter far from the centre of galaxies does not follow the inverse square dependence. There could be many reasons for this effect, and a change in the equations for large distances might be one of them. This issue is still open.

Theoreticians have explored many alternative equations, such as scalar-tensor theories, theories with torsion, or theories which break Lorentz invariance. However, none of these seem to fit experimental data yet.

The limits of general relativity

Even though successful, the description of motion presented so far is unsatisfactory; Challenge 883 ny maybe you already have some gut feeling about certain unresolved issues. First of all, even though the speed of light is the starting point of the whole theory, we still do not know what light actually *is*. Finding out will be our next topic.

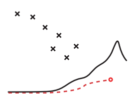
Secondly, we saw that everything falls along geodesics. But a mountain does not fall. Somehow the matter below prevents it from falling. How does it achieve this? And where does mass come from anyway? What is mass? General relativity does not provide an answer; in fact, it does not describe matter *at all*. Einstein used to say that the left-hand side of the field equations, describing the curvature of space-time, was granite, the right-hand side, describing matter, was sand. Indeed, at this point we still do not know what mass is. As already remarked, to change the sand into rock we first need quantum theory and then, in a further step, its unification with relativity. This is also the program for the rest of our adventure.

We also saw that matter is necessary to clearly distinguish space and time, and in particular, to understand the working of clocks, meter bars and balances. In particular, one question remains: why are there units of mass, length and time in nature *at all*? This deep question will also be addressed in the following.

Additionally, we found how little we know about the vacuum. We need to understand the magnitude of the cosmological constant and the number of space-time dimensions. Only then can we answer the simple question: Why is the sky so far away? General relativity does not help here. Worse, the smallness of the cosmological constant contradicts the simplest version of quantum theory; this is one of the reasons why we still have quite Page 918 some height to escalate before we reach the top of Motion Mountain.

In short, to describe motion well, we realize that we need a more precise description of light, of matter and of the vacuum. In short, we need to know more about everything Page 783 we know. Otherwise we cannot hope to answer questions about mountains, clocks and stars. In a sense, it seems that we achieved quite little. Fortunately, this is not true. We learned so much that for the following topic we are forced to go backwards, to situations *without* gravity, i.e. back to the framework of special relativity. That is the next, middle section of our mountain ascent. Despite the simplification to flat space-time, a lot of fun is waiting there.

> It's a good thing we have gravity, or else when birds died they'd just stay right up there. Hunters would be all confused.
>
> Steven Wright, comedian.



Bibliography

A man will turn over half a library to make one book.
Samuel Johnson (1709–1784)

301 The simplest historical source is ALBERT EINSTEIN, *Sitzungsberichte der Preussischen Akademie der Wissenschaften* II pp. 844–846, 1915. It is the first explanation of the general theory of relativity, in only three pages. The theory is then explained in detail in the famous article ALBERT EINSTEIN, Die Grundlage der allgemeinen Relativitätstheorie, *Annalen der Physik* **49**, pp. 769–822, 1916. The historic references can be found in German and English language in JOHN STACHEL, ed., *The Collected Papers of Albert Einstein*, Volumes 1–9, Princeton University Press, 1987–.

A selection of English language textbooks for deeper study should include, in ascending order of depth and difficulty:

▪ An entertaining book without any formula, but nevertheless accurate and detailed, is the paperback by IGOR NOVIKOV, *Black Holes and the Universe*, Cambridge University Press, 1990.

▪ Almost no formulae, but loads of insight are found in the enthusiastic text by JOHN A. WHEELER, *A Journey into Gravity and Space-Time*, W.H. Freeman, 1990.

▪ An excellent didactical presentation is EDWIN F. TAYLOR & JOHN A. WHEELER, *Exploring Black Holes: Introduction to General Relativity*, Addison Wesley Longman 2000.

▪ Beauty, simplicity and shortness is the characteristic of MALCOLM LUDVIGSEN, *General Relativity, a Geometric Approach*, Cambridge University Press, 1999.

▪ Good explanation is the strength of BERNARD SCHUTZ, *Gravity From the Ground Up*, Cambridge University Press, 2003.

▪ A good overview of experiments and theory is given in JAMES FOSTER & J.D. NIGHTINGALE, *A Short Course in General Relativity*, Springer Verlag, 2nd edition 1998.

▪ A pretty text is SAM LILLEY, *Discovering Relativity for Yourself*, Cambridge University Press, 1981.

▪ A modern text is by RAY D'INVERNO *Introducing Einstein's Relativity*, Clarendon Press, 1992. It includes an extended description of black holes and gravitational radiation, and regularly refers to present research.

▪ A beautiful, informative and highly recommended text is HANS C. OHANIAN & REMO RUFFINI, *Gravitation and Spacetime*, W.W. Norton & Co., 1994.

▪ A well written and modern book, with emphasis on the theory, by one of the great masters of the field is WOLFGANG RINDLER, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001.

▪ A classic is STEVEN WEINBERG, *Gravitation and Cosmology*, Wiley, 1972.

▪ The passion of general relativity can be experienced also in JOHN KLAUDER, editor, *Magic without Magic: John Archibald Wheeler – A Collection of Essays in Honour of His Sixtieth Birthday*, W.H. Freeman & Co., San Francisco, 1972.

▪ An extensive text is KIP S. THORNE, *Black Holes and Time Warps – Einstein's Outrageous Legacy*, W.W. Norton, 1994.

▪ The most mathematical – and toughest – text is ROBERT M. WALD, *General Relativity*, University of Chicago Press, 1984.

For texts in other languages, see the next reference. In fact, it is time for a large and modern textbook on general relativity.

▪ Much information about general relativity is available on the internet. As a good starting point for US-American material, see the http://math.ucr.edu/home/baez/relativity.html website. Cited on pages 344, 376, and 377.

302 A beautiful German teaching text is the classic G. Falk & W. Ruppel, Mechanik, Relativität, Gravitation – ein Lehrbuch, Springer Verlag, Dritte Auflage, 1983.

A practical and elegant booklet is Ulrich E. Schröder, *Gravitation – Einführung in die allgemeine Relativitätstheorie*, Verlag Harri Deutsch, Frankfurt am Main, 2001.

A modern reference is Torsten Fliessbach, *Allgemeine Relativitätstheorie*, Akademischer Spektrum Verlag, 1998.

Excellent is Hubert Goenner, *Einführung in die spezielle und allgemeine Relativitätstheorie*, Akademischer Spektrum Verlag, 1996.

In Italian, there is the beautiful, informative, but expensive Hans C. Ohanian & Remo Ruffini, *Gravitazione e spazio-tempo*, Zanichelli, 1997. It is highly recommended. Cited on pages 344, 373, 374, 376, 377, and 469.

303 P. Mohazzabi & J.H. Shea, High altitude free fall, *American Journal of Physics* **64**, pp. 1242–1246, 1996. As a note, due to a technical failure Kittinger had his hand in (near) vacuum during his ascent, without incurring any permanent damage. On the consequences of human exposure to vacuum, see the http://www.sff.net/people/geoffrey.landis/vacuum.html website. Cited on page 344.

304 This story is told e.g. by W.G. Unruh, Time, gravity, and quantum mechanics, preprint available at http://www.arxiv.org/abs/gr-qc/9312027. Cited on page 345.

305 H. Bondi, Gravitation, *European Journal of Physics* **14**, pp. 1–6, 1993. Cited on page 346.

306 J.W. Brault, Princeton University PhD thesis, 1962. See also J.L. Snider, *Physical Review Letters* **28**, pp. 853–856, 1972, and for the star Sirius see J.L. Greenstein & al., *Astrophysical Journal* **169**, p. 563, 1971. Cited on pages 346 and 462.

307 R.V. Pound & G.A. Rebka, The famous paper is Apparent weight of photons, *Physical Review Letters* **4**, pp. 337–341, 1960. A higher precision version was published by R.V. Pound & J.L. Snider, *Physical Review Letters* **13**, p. 539, 1964, and R.V. Pound & J.L. Snider, *Physical Review B* **140**, p. 788, 1965, Cited on pages 347 and 462.

308 J.C. Hafele & Richard E. Keating, Around-the-world atomic clocks: predicted relativistic time gains, *Science* **177**, pp. 166–167, and Around-the-world atomic clocks: observed relativistic time gains, pp. 168–170, 14 July 1972. Cited on page 347.

309 R.F.C. Vessot & al., Test of relativistic gravitation with a space-borne hyrogen maser, *Physical Review Letters* **45**, pp. 2081–2084, 1980. The experiment was performed in 1976; there are more than a dozen coauthors involved in this work, which involved shooting a maser into space with a scout missile to a height of *c.* 10 000 km. Cited on page 347.

310 L. Briatore & S. Leschiutta, Evidence for the Earth gravitational shift by direct atomic-time-scale comparison, *Il Nuovo Cimento* **37B**, pp. 219–231, 1977. Cited on page 347.

311 More information about tides can be found in E.P. Clancy, *The Tides*, Doubleday, New York, 1969. Cited on page 348.

312 The expeditions had gone to two small islands, namely to Sobral, north of Brazil, and to Principe, in the gulf of Guinea. The results of the expedition appeared in the Times newspaper *before* they appeared in a scientific journal. Today this would be seen as a gross violation of scientific honesty. The results were published as F.W. Dyson, A.S. Eddington & C. Davidson, *Philosophical Transactions of the Royal Society (London)* **220A**, p. 291, 1920, and *Memoirs of the Royal Astronomical Society* **62**, p. 291, 1920. Cited on page 349.

313 A good source for images of space-time is the text by G.F.R. Ellis & R. Williams, *Flat and Curved Space-times*, Clarendon Press, Oxford, 1988. Cited on page 350.

314 J. Droste, Het veld van een enkel centrum in Einstein's theorie der zwaartekracht, en de beweging van een stoffelijk punt, *Verslag gew. Vergad. Wiss. Amsterdam* **25**, pp. 163–180, 1916. Cited on page 351.

315 The name black hole was introduced in 1967 at a pulsar conference, as described in his autobiography by John A. Wheeler, *Geons, Black Holes, and Quantum Foam: A Life in Physics*, W.W. Norton, 1998, pp. 296–297: 'In my talk, I argued that we should consider the possibility that at the center of a pulsar is a gravitationally completely collapsed object. I remarked that one couldn't keep saying "gravitationally completely collapsed object" over and over. One needed a shorter descriptive phrase. "How about black hole?" asked someone in the audience. I had been searching for just the right term for months, mulling it over in bed, in the bathtub, in my car, whenever I had quiet moments. Suddenly, this name seemed exactly right. When I gave a more formal ...lecture ...a few weeks later on, on December 29, 1967, I used the term, and then included it into the written version of the lecture published in the spring of 1968 ...I decided to be casual about the term "black hole", dropping it into the lecture and the written version as if it were an old familiar friend. Would it catch on? Indeed it did. By now every schoolchild has heard the term.'

The widespread usage of the term was started by the article by R. Ruffini & J.A. Wheeler, Introducing the black hole, *Physics Today* **24**, pp. 30–41, January 1971.

In his autobiography, Wheeler also writes that the expression 'black hole has no hair' was criticized as 'obscene' by Feynman. An interesting comment by a physicist who used to write his papers in topless bars. Cited on pages 352, 441, and 446.

316 L.B. Kreuzer, Experimental measurement of the equivalence of active and passive gravitational mass, *Physical Review* **169**, pp. 1007–1012, 1968. With a clever set-up, he showed that the gravitational mass of fluorine and of bromine are equal. Cited on page 352.

317 G.W. Gibbons, The maximum tension principle in general relativity, *Foundations of Physics* **32**, pp. 1891–1901, 2002, or http://www.arxiv.org/abs/hep-th/0210109. Cited on page 354.

318 That bodies fall along geodesics has been checked by ... Cited on page 356.

319 So far, the experiments confirm that electrostatic and (strong) nuclear energy fall like matter to within one part in 10^8, and weak (nuclear) energy to within a few per cent. This is summarized in the reference Ref. 324. Cited on page 356.

320 J. Soldner, *Berliner Astronomisches Jahrbuch auf das Jahr 1804*, 1801, p. 161. Cited on page 357.

321 See the excellent explanation in N. Dragon, Geometrie der Relativitätstheorie, downloadable from the http://www.itp.uni-hannover.de/~dragon website. No citations.

322 See for example K.D. Olum, Superluminal travel requires negative energies, *Physical Review Letters* **81**, pp. 3567–3570, 1998, or M. Alcubierre, The warp drive: hyper-fast travel within general relativity, *Classical and Quantum Gravity* **11**, pp. L73–L77, 1994. See also Chris Van Den Broeck, A warp drive with more reasonable total energy requirements, *Classical and Quantum Gravity* **16**, pp. 3973–3979, 1999. Cited on page 359.

323 See the *Astronomical Almanac*, and its *Explanatory Supplement*, H.M. Printing Office, London and U.S. Government Printing Office, Washington, 1992. For the information about various time coordinates used in the world, such as barycentric coordinate time, the time at the barycentre of the solar system, see also the http://tycho.usno.navy.mil/systime.html web page. It also contains a good bibliography. Cited on page 359.

324 An overview is given in C. Will, *Theory and Experiment in Gravitational Physics*, chapter 14.3, revised edition, Cambridge University Press, 1993. (Despite being a standard reference, his view the role of tides and the role of gravitational energy within the principle of equivalence has been criticised by other researchers.) See also C. Will, *Was Einstein Right? –*

Putting General Relativity to the Test, Oxford University Press, 1993. See also his paper http://www.arxiv.org/abs/gr-qc/9811036. Cited on pages 360, 376, and 467.

325 The calculation omits several smaller effects, such as rotation of the Earth, red-shift and others. For the main effect, see Edwin F. Taylor, 'The boundaries of nature: special and general relativity and quantum mechanics, a second course in physics' – Edwin F. Taylor's acceptance speech for the 1998 Oersted Medal presented by the American Association of Physics Teachers, 6 January 1998, *American Journal of Physics* **66**, pp. 369–376, 1998. Cited on page 360.

326 A.G. Lindh, Did Popper solve Hume's problem?, *Nature* **366**, pp. 105–106, 11 November 1993, Cited on page 360.

327 The measurement is presented in the Astrophysical Journal, in 1998 or 1999. Some beautiful graphics at the http://www.physics.uiuc.edu/groups/tastro/movies/spm/ website show the models of this star system. Cited on page 361.

328 R.J. Nemiroff, Visual distortions near a black hole and a neutron star, *American Journal of Physics* **61**, pp. 619–632, 1993. Cited on page 361.

329 The equality was first tested with precision by Roland von Eőtvős, *Annalen der Physik & Chemie* **59**, p. 354, 1896, and by R. von Eőtvős, V. Pekár & E. Fekete, Beiträge zum Gesetz der Proportionalität von Trägheit und Gravität, *Annalen der Physik 4, Leipzig* **68**, pp. 11–66, 1922. He found agreement to 5 parts in 10^9. More experiments were performed by P.G. Roll, R. Krotkow & R.H. Dicke, The equivalence of inertial and passive gravitational mass, *Annals of Physics (NY)* **26**, pp. 442–517, 1964, one of the most interesting and entertaining research articles in experimental physics, and by V.B. Braginsky & V.I. Panov, *Soviet Physics – JETP* **34**, pp. 463–466, 1971. Modern results, with errors less than one part in 10^{12}, are by Y. Su & al., New tests of the universality of free fall, *Physical Review* **D50**, pp. 3614–3636, 1994. Several experiments have been proposed to test the equality in space to less than one part in 10^{16}. Cited on pages 361 and 462.

330 The Thirring effect was predicted in H. Thirring, Über die Wirkung rotierender ferner Massen in der Einsteinschen Gravitationstheorie, *Physikalische Zeitschrift* **19**, pp. 33–39, 1918, and in H. Thirring, Berichtigung zu meiner Arbeit: "Über die Wirkung rotierender Massen in der Einsteinschen Gravitationstheorie", *Physikalische Zeitschrift* **22**, p. 29, 1921. The Lense–Thirring effect was predicted in J. Lense & H. Thirring, Über den Einfluß der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie, *Physikalische Zeitschrift* **19**, pp. 156–163, 1918. See also reference Ref. 348. Cited on page 363.

331 The feat used the LAGEOS and LAGEOS II satellites and is told in Ignazio Ciufolini, The 1995-99 measurements of the Thirring–Lense effect using laser-ranged satellites, *Classical and Quantum Gravity* **17**, pp. 2369–2380, 2000. See also I. Ciufolini & E.C. Pavlis, A confirmation of the general relativistic prediction of the Lense–Thirring effect, *Nature* **431**, pp. 958–960, 2004. Cited on pages 364, 367, and 462.

332 The detection of the Lense–Thirring effect in binary pulsars is presented in R.D. Blandford, Lense-Thirring precession of radio pulsars, *Journal of Astrophysics and Astronomy* **16**, pp. 191–206, 1995. Cited on page 364.

333 G. Holzmüller, *Zeitschrift für Mathematik und Physik* **15**, p. 69, 1870, F. Tisserand, *Comptes Rendus* **75**, p. 760, 1872, and *Comptes Rendus* **110**, p. 313, 1890. Cited on page 365.

334 B. Mashhoon, Gravitoelectromagnetism, http://www.arxiv.org/abs/gr-qc/0011014. See also its extensive reference list on gravitomagnetism. Cited on page 365.

335 D. Bedford & P. Krumm, On relativistic gravitation, *American Journal of Physics* **53**,

pp. 889–890, 1985, and P. KRUMM & D. BEDFORD, The gravitational Poynting vector and energy transfer, *American Journal of Physics* 55, pp. 362–363, 1987. Cited on pages 366 and 372.

336 This is told in JOHN A. WHEELER, *A Journey into Gravity and Space-Time*, W.H. Freeman, 1990. Cited on page 368.

337 See for example, K.T. MCDONALD, Answer to question #49. Why c for gravitational waves?, *American Journal of Physics* **65**, pp. 591–592, 1997, and section III of V.B. BRAGINSKY, C.M. CAVES & K.S. THORNE, Laboratory experiments to test relativistic gravity, *Physical Review D* **15**, pp. 2047–2068, 1992. Cited on page 369.

338 The original claim is by S.M. KOPEIKIN, The post-Newtonian treatment of the VLBI experiment on September 8, 2002, *Physics Letters A* **312**, pp. 147–157, 2003, or http://www.arxiv.org/abs/gr-qc/0212121. An argument against the claim was published, among others, by STUART SAMUEL, On the speed of gravity and the v,c corrections to the Shapiro time delay/ http://www.arxiv.org/abs/astro-ph/0304006. Cited on pages 370 and 374.

339 The quadrupole formula is explained in the text by Goenner. See reference Ref. 302. Cited on page 372.

340 For an introduction to gravitational waves, see B.F. SCHUTZ, *American Journal of Physics* **52**, p. 412, 1997, No citations.

341 The beautiful summary by DANIEL KLEPPNER, The gem of general relativity, *Physics Today* pp. 9–11, April 1993, appeared half a year before the authors of the cited work, Joseph Taylor and Russel Hulse, received the Nobel prize for the discovery of millisecond pulsars. A more detailed review article is J.H. TAYLOR, Pulsar timing and relativistic gravity, *Philosophical Transactions of the Royal Society, London A* **341**, pp. 117–134, 1992. The original paper is J.H. TAYLOR & J.M. WEISBERG, Further experimental tests of relativistic gravity using the binary pulsar PSR 1913+16, *Astrophysical Journal* **345**, pp. 434–450, 1989. See also J.M. WEISBERG, J.H. TAYLOR & L.A. FOWLER, Pulsar PSR 1913+16 sendet Gravitationswellen, *Spektrum der Wissenschaft*, Dezember 1981. Cited on pages 373 and 374.

342 W.B. BONNOR & M.S. PIPER, The gravitational wave rocket, *Classical and Quantum Gravity* **14**, pp. 2895–2904, 1997, or http://www.arxiv.org/abs/gr-qc/9702005. Cited on page 374.

343 L. LERNER, A simple calculation of the deflection of light in a Schwarzschild gravitational field, *American Journal of Physics* **65**, pp. 1194–1196, 1997. Cited on page 375.

344 A. EINSTEIN, Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes, *Annalen der Physik* **35**, p. 898, 1911. Cited on page 376.

345 I.I. SHAPIRO & al., Fourth test of general relativity, *Physical Review Letters* **13**, pp. 789–792, 1964. Cited on page 376.

346 I.I. SHAPIRO & al., Fourth test of general relativity: preliminary results, *Physical Review Letters* **20**, pp. 1265–1269, 1968. Cited on page 376.

347 J.H. TAYLOR, Pulsar timing and relativistic gravity, *Proceedings of the Royal Society, London A* **341**, pp. 117–134, 1992. Cited on pages 376 and 378.

348 W. DE SITTER, On Einstein's theory of gravitation and its astronomical consequences, *Monthly Notes of the Royal Astrononmical Society* **77**, pp. 155–184, p. 418E, 1916. For a discussion of de Sitter precession and Lense–Thirring precession, see also B.R. HOLSTEIN, Gyroscope precession in general relativity, *American Journal of Physics* **69**, pp. 1248–1256, 2001. Cited on pages 379 and 468.

349 B. BERTOTTI, I. CIUFOLINI & P.L. BENDER, New test of general relativity: measurement of de Sitter geodetic precession rate for lunar perigee, *Physical Review Letters* **58**, pp.

1062–1065, 1987. Later it was confirmed by I.I. Shapiro & al., Measurement of the de Sitter precession of the moon: a relativistic three body effect, *Physical Review Letters* **61**, pp. 2643–2646, 1988. Cited on pages 379 and 462.

350 Wolfgang Rindler, *Essential Relativity*, revised second edition, Springer, 1977. Cited on page 381.

351 This is told (without the riddle solution) on p. 67, in Wolfgang Pauli, *Relativitätstheorie*, Springer Verlag, Berlin, 2000, the edited reprint of a famous text originally published in 1921. The reference is H. Vermeil, Notiz über das mittlere Krümmungsmaß einer n-fach ausgedehnten Riemannschen Mannigfaltkigkeit, *Göttinger Nachrichten, mathematische–physikalische Klasse* p. 334, 1917. Cited on page 381.

352 M. Santander, L.M. Nieto & N.A. Cordero, A curvature based derivation of the Schwarzschild metric, *American Journal of Physics* **65**, pp. 1200–1209, 1997. Cited on pages 384 and 386.

353 Richard P. Feynman, Fernando B. Morinigo, William G. Wagner & Brian Hatfield, *Feynman Lectures on Gravitation*, Westview Press, 1995. Cited on page 385.

354 C.G. Torre & I.M. Anderson, Symmetries of the Einstein Equations, *Physical Review Letters* **70**, pp. 3525–3529, 1993, or http://www.arxiv.org/abs/gr-qc/9302033. Cited on page 391.

355 Y. Wang & M. Tegmark, New dark energy constraints from supernovae, microwave background and galaxy clustering, *Physical Review Letters* **92**, p. 241302, 2004, or http://www.arxiv.org/astro-ph/0403292. Cited on page 392.

356 Arguments for the emptyness of general covariance are given by John D. Norton, General covariance and the foundations of general relativity, *Reports on Progress in Physics* **56**, pp. 791–858, 1993. The opposite point, including the discussion of 'absolute elements,' is made in the book by J.L. Anderson, *Principles of Relativity Physics*, chapter 4, Academic Press, 1967. Cited on page 392.

357 See for example R.A. Knop & al., New constraints on Omega_M, Omega_Lambda, and w from an independent set of eleven high-redshift supernovae observed with HST, *Astrophysical Journal* **598**, pp. 102–137, 2003. Cited on page 394.

358 R.P. Feynman, R.B. Leighton & M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, 1977, volume II, p. 42-14. Cited on page 396.

359 A recent overview on the experimental tests of the universality of free fall is given by R.J. Hughes, The equivalence principle, *Contemporary Physics* **4**, pp. 177–191, 1993. Cited on page 397.

360 See for example H.L. Bray, Black holes, geometric flows, and the Penrose inequality in general relativity, *Notices of the AMS* **49**, pp. 1372–1381, 2002. Cited on page 398.

361 See for example the paper by K. Dalton, Gravity, Geometry and Equivalence, preprint to be found at http://www.arxiv.org/abs/gr-qc/9601004 and L. Landau & E. Lifshitz, The Classical Theory of Fields, Pergamon, 4th edition, 1975, p. 241. Cited on page 399.

362 Ekkehart Kröner, *Kontinuumstheorie der Versetzungen und Eigenspannungen*, Springer, 1958. Kröner shows how to use the Ricci formalism in the solid state. Cited on page 400.

363 The equivalence of the various definitions of the Riemann tensor is explained in ... Cited on page 401.

364 Black Cited on page 400.

365 This famous quote is the first sentence of the final chapter, the 'Beschluß', of Immanuel Kant, *Kritik der praktischen Vernunft*, 1797. Cited on page 402.

366 Aetius, Opinions, III, I, 6. See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 445, 1991. Cited on page 402.

367 Galaxy mass measurements are described by ... Cited on page 404.

368 See for example A.N. Cox, editor *Allen's Astrophysical Quantities*, AIP Press and Springer Verlag, 2000. Cited on page 406.

369 P. Jetzer, Gravitational microlensing, *Naturwissenschaften* **86**, pp. 201–211, 1999. Cited on page 407.

370 D. Figer, An upper limit to the masses of stars, *Nature* **434**, pp. 192–194, 2005. Cited on page 408.

371 G. Basri, The discovery of brown dwarfs, *Scientific American* pp. 77–83, April 2001. Cited on page 408.

372 P.M. Woods & C. Thompson, Soft gamma repeaters and anomalous X-ray pulsars: magnetar candidates, http://www.arxiv.org/abs/astro-ph/0406133. Cited on page 408.

373 B.M. Gaensler, N.M. McClure-Griffiths, M.S. Oey, M. Haverkorn, J.M. Dickey & A.J. Green , A stellar wind bubble coincident with the anomalous X-ray pulsar 1E 1048.1-5937: are magnetars formed from massive progenitors?, *The Astrophysical Journal (Letters)* **620**, pp. L95–L98, 2005, or http://www.arxiv.org/abs/astro-ph/0501563. Cited on page 408.

374 An opposite idea is defended by ... Cited on page 412.

375 C. Wirtz, *Scientia* **38**, p. 303, 1925, and K. Lundmark, The motions and the distances of the spiral nebulae, *Monthly Notices of the Royal Astronomical Society* **85**, p. 865-894, 1925. See also G. Stromberg, Analysis of radial velocities of globular clusters and non-galactic nebulae, *Astrophysical Journal* **61**, pp. 353–362, 1925. Cited on page 412.

376 G. Gamow, The origin of the elements and the separation of galaxies, *Physical Review* **74**, p. 505, 1948. Cited on page 413.

377 A.G. Doroshkevich & I.D. Novikov, *Dokl. Akad. Nauk. SSSR* **154**, p. 809, 1964. It appeared translated into English a few months later. The story of the prediction was told by Penzias in his Nobel lecture. Cited on page 413.

378 Arno A. Penzias & Robert W. Wilson, A measurement of excess antenna temperature at 4080 Mcs, *Astrophysical Journal* **142**, pp. 419–421, 1965. Cited on page 413.

379 Macrobus, Dream of Scipio, XIV, 19. See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 61, 1991. Cited on page 415.

380 The first oxygen seems to have appeared in the atmosphere, produced by microorganisms, 2.32 thousand million years ago. See A. Becker & al., Dating the rise of atmospheric oxygen, *Nature* **427**, p. 117-120, 2003. Cited on page 416.

381 Gabriele Walker, *Snowball Earth – The Story of the Great Global Catastrophe That Spawned Life as We Know It*, Crown Publishing 2003. Cited on pages 416 and 767.

382 A. Friedmann, *Zeitschrift für Physik* **10**, p. 377, 1922, and *Zeitschrift für Physik* **21**, p. 326, 1924. Cited on page 418.

383 H. Knutsen, Darkness at night, *European Journal of Physics* **18**, pp. 295–302, 1997. Cited on pages 423 and 424.

384 See for example P.D. Peşić, Brightness at night, *American Journal of Physics* **66**, pp. 1013–1015, 1998. Cited on pages 423 and 424.

385 Paul Wesson, Olbers' paradox and the spectral intensity of extra-galactic background light, *Astrophysical Journal* **367**, p. 399, 1991. Cited on page 424.

386 Steven Weinberg, *Gravitation and Cosmology*, John Wiley, 1972. An excellent book written with a strong personal touch and stressing most of all the relation with experimental data. It does not develop a strong feeling for space-time curvature, and does not address the basic problems of space and time in general relativity. Excellent for learning how to actually calculate things, but less for the aims of our mountain ascent. Cited on pages 424 and 455.

387 Supernova searches are being performed by ... Cited on page 425.

388 The experiments are discussed in detail in the excellent review by D. Giulini & N. Straumann, Das Rätsel der kosmischen Vakuumenergiedichte und die beschleunigte Expansion des Universums, *Physikalische Blätter* **556**, pp. 41–48, 2000. See also N. Straumann, The mystery of the cosmic vacuum energy density and the accelerated expansion of the universe, *European Journal of Physics* **20**, pp. 419–427, 1999. Cited on pages 425 and 462.

389 A. Harvey & E. Schucking, Einstein's mistake and the cosmological contant, *American Journal of Physics* **68**, pp. 723–727, 2000. Cited on page 425.

390 The author of the bible explains *rain* in this way, as can be deduced from its very first page, Genesis 1, 6-7. Cited on page 426.

391 Up to his death, Fred Hoyle defended his belief that the universe is static, e.g. in G. Burbidge, F. Hoyle & J.V. Narlikar, A different approach to cosmology, *Physics Today* **52**, pp. 38–44, 1999. The team has also written a book with the same title, published in 2000 by Cambridge University Press. Cited on page 427.

392 Stephen W. Hawking & G.F.R. Ellis, The large scale structure of space-time, Cambridge University Press, Cambridge, 1973. Among others, this reference text discusses the singularities of space-time, and their necessity in the history of the universe. Cited on pages 427, 457, and 475.

393 Augustine, *Confessions*, 398, writes: 'My answer to those who ask 'What was god doing before he made heaven and Earth?' is not 'He was preparing hell for people who pry into mysteries'. This frivolous retort has been made before now, so we are told, in order to evade the point of the question. But it is one thing to make fun of the questioner and another to find the answer. So I shall refrain from giving this reply. [...] But if before heaven and Earth there was no time, why is it demanded what you [god] did then? For there was no "then" when there was no time.' (Book XI, chapter 12 and 15) Cited on page 428.

394 Stephen Hawking, *A Brief History of Time – From the Big Bang to Black Holes*, 1988. Reading this bestseller is almost a must for any physicist, as it is a frequent topic at dinner parties. Cited on page 428.

395 Star details are explained in many texts. See for example ... Cited on page 430.

396 J. Pelt, R. Kayser, S. Refsdal & T. Schramm, The light curve and the time delay of QSO 0957+561, *Astronomy and Astrophysics* **305**, p. 97, 1996. Cited on page 432.

397 F. Zwicky, *Physical Review Letters* **51**, p. 290, and p. 679, 1937. The negative view by Einstein is found in A. Einstein, *Science* **84**, p. 506, 1936. A review on gravitational lensing can even be found online, in the paper by J. Wambsganss, *Living Reviews in Relativity* **1-12**, pp. 1–80, 1998, to be found on the http://www.livingreviews.org/Articles/Volume1/1998-12wamb/ website.

There is also the book by P. Schneider, J. Ehlers & E.E. Falco, *Gravitational Lenses*, Springer Verlag, Berlin, 1992. Cited on page 432.

398 M. Lachièze-Rey & J.-P. Luminet, Cosmic topology, *Physics Reports* **254**, pp. 135–214, 1995. See also B.F. Roukema, Cosmic topology, http://www.arxiv.org/abs/astro-ph/

0010185 preprint. Cited on page 434.

399 Thanks to Steve Carlip for clarifying this point. Cited on page 434.

400 G.F.R. Ellis & T. Rothmann, Lost horizons, *American Journal of Physics* **61**, pp. 883–893, 1993. Cited on page 435.

401 A. Guth, *Die Geburt des Kosmos aus dem Nichts – Die Theorie des inflationären Universums*, Droemer Knaur, München 1999. Cited on page 435.

402 Entropy values for the universe have been discussed by Ilya Prigogine, *Is Future Given?*, World Scientific, Singapore 2003. This was his last book. For a different approach, see G.A. Mena Marugán & S. Carneiro, Holography and the large number hypothesis, electronic preprint at http://www.arxiv.org/abs/gr-qc/0111034. It also repeats the often heard statement that the universe has an entropy that is much smaller than the theoretical maximum. The maximum is 'generally' estimated to be $10^{120}\,k$, whereas the actual value is 'estimated' to be $10^{100}\,k$. However, other authors give $10^{84}\,k$.

Challenge 884 ny

In 1974, Roger Penrose also made statements about the entropy of the universe. Cited on page 436.

403 C.L. Bennet, M.S. Turner & M. White, The cosmic rosetta stone, *Physics Today* **50**, pp. 32–38, November 1997. The cosmic background radiation differs from black hole radiation by less than 0.005 %. Cited on page 437.

404 The lack of expansion in the solar system is shown in ... Cited on page 438.

405 A pretty article explaining how one can make experiments to find out how the human body senses rotation even when blindfolded and earphoned is described by M.-L. Mittelstaedt & H. Mittelstaedt, The effect of centrifugal force on the perception of rotation about a vertical axis, *Naturwissenschaften* **84**, pp. 366–369, 1997. Cited on page 438.

406 The independence of inertia has been tested ... Cited on page 438.

407 The present status is given in the conference proceedings by Julian Barbour & Herbert Pfister, (editors) Mach's principle: from Newton's bucket to quantum gravity, Birkhäuser, 1995. Various formulations of Mach's principle – in fact, 21 different ones – are compared on page 530.

In a related development, in 1953, Dennis Sciama published a frequently cited paper in which he argues that inertia of a particle is due to the gravitational attraction of all other matter in the universe. The paper is widely quoted, but makes no new statements on the issue. See D.W. Sciama, On the origin of inertia, *Monthly Notices of the Royal Astronomical Society* **113**, pp. 34–42, 1953. Cited on pages 438 and 439.

408 Information on the rotation of the universe is given in A. Kogut, G. Hinshaw & A.J. Banday, Limits to global rotation and shear from the COBE DMR four-year sky maps, *Physical Review D* **55**, pp. 1901–1905, 1997. Earlier information is found in J.D. Barrow, R. Juszkiewicz & D.H. Sonoda, Universal rotation: how large can it be?, *Monthly Notices of the Royal Astronomical Society* **213**, pp. 917–943, 1985. See also J.D. Barrow, R. Juszkiewicz & D.H. Sonoda, Structure of the cosmic microwave background, *Nature* **309**, pp. 397–402, 1983, or E.F. Bunn, P.G. Fereira & J. Silk, How anisotropic is the universe?, *Physical Review Letters* **77**, pp. 2883–2886, 1996. Cited on page 439.

409 The issue has been discussed within linearized gravity by Richard Tolman, in his textbook *Relativity, Thermodynamics, and Cosmology*, Clarendon Press, 1934, on pp. 272–290. The exact problem has been solved by A. Peres, Null electromagnetic fields in general relativity theory, *Physical Review* **118**, pp. 1105–1110, 1960, and by W.B. Bonnor, The gravitational field of light, *Commun. Math. Phys.* **13**, pp. 163–174, 1969. See also N.V. Mitskievic

& K.K. Kumaradtya, The gravitational field of a spinning pencil of light, *Journal of Mathematical Physics* **30**, pp. 1095–1099, 1989, and P.C. Aichelburg & R.U. Sexl, On the gravitational field of a spinning particle, *General Relativity and Gravitation* **2**, pp. 303–312, 1971. Cited on page 440.

410 See the delightful popular account by Igor Novikov, *Black Holes and the Universe*, Cambridge University Press 1990. The consequences of light decay were studied by M. Bronstein, Die Ausdehnung des Weltalls, *Physikalische Zeitschrift der Sowjetunion* **3**, pp. 73–82, 1933. Cited on pages 440 and 444.

411 C.L. Carilli, K.M. Menten, J.T. Stocke, E. Perlman, R. Vermeulen, F. Briggs, A.G. de Bruyn, J. Conway & C.P. Moore, Astronomical constraints on the cosmic evolution of the fine structure constant and possible quantum dimensions, *Physical Review Letters* **85**, pp. 5511–5514, 25 December 2000. Cited on page 440.

412 An excellent and entertaining book on black holes, without any formula, but nevertheless accurate and detailed, is the paperback by Igor Novikov, *Black Holes and the Universe*, Cambridge University Press, 1990. See also Edwin F. Taylor & John A. Wheeler, *Exploring Black Holes: Introduction to General Relativity*, Addison Wesley Longman 2000.

For a historical introduction, see the paper by R. Ruffini, The physics of gravitationally collapsed objects, pp. 59–118, in *Neutron stars, black holes and binary X-ray sources*, Proceedings of the Annual Meeting, San Francisco, Calif., February 28, 1974, Reidel Publishing 1975. Cited on page 441.

413 The observations of black holes at the centre of galaxies and elsewhere are summarised by R. Blandford & N. Gehrels, Revisiting the black hole, *Physics Today* **52**, June 1999. Cited on pages 441 and 452.

414 J. Michell, *Philosophical Transactions of the Royal Society London* **74**, p. 35, 1784, reprinted in S. Detweiler, *Black Holes – Selected Reprints*, American Association of Physics Teachers, 1982. Cited on page 441.

415 The beautiful paper is R. Oppenheimer & H. Snyder, On continued gravitational contraction, *Physical Review* **56**, pp. 455–459, 1939. Cited on page 442.

416 R.P. Kerr, Gravitational field of a spinning mass as an example of algebraically special metrics, *Physical Review Letters* **11**, pp. 237–238, 1963. Cited on page 446.

417 E.T. Newman, E. Couch, R. Chinnapared, A. Exton, A. Prakash & R. Torrence, Metric of a rotating, charged mass, *Journal of Mathematical Physics* **6**, pp. 918–919, 1965. Cited on page 446.

418 For a summary, see P.O. Mazur, Black hole uniqueness theorems, pp. 130–157, in M.A.H. Mac Callum, editor, *General Relativity and Gravitation*, 1987. Cited on page 446.

419 H.P. Künzle & A.K.M. Masood-ul-Alam, Spherically symmetric static SU(2) Einstein-Yang-Mills fields, *Journal of Mathematical Physics* **31**, pp. 928–935, 1990. Cited on page 446.

420 About the tendency of gravitational radiation to produce spherical shapes, see for example ... Cited on page 447.

421 R. Penrose & R.M. Floyd, Extraction of rotational energy from a black hole, *Nature* **229**, pp. 177–179, 1971. Cited on page 448.

422 The mass energy relation for a rotating black hole is due to D. Christodoulou, Reversible and irreversible transformations in black hole physics, *Physical Review Letters* **25**, pp. 1596–1597, 1970. For a general, charged and rotating black hole it is due to D. Christodoulou & R. Ruffini, Reversible transformations of a charged black hole, *Physical Review D* **4**, pp. 3552–3555, 1971. Cited on page 448.

423 J.D. Bekenstein, Black holes and entropy, *Physical Review* **D7**, pp. 2333–2346, 1973. Cited on page 449.

424 The paradox is discussed in M.A. Abramowicz, Black hole and the centrifugal force paradox, *Scientific American* pp. 23–31, March 1993, and in the comment by Don N. Page, Relative alternatives, *Scientific American* p. 5, August 1993. See also M.A. Abramowicz & E. Szuszkiewicz, The wall of death, *American Journal of Physics* **61**, pp. 982–991, 1993, and M.A. Abramowicz & J.P. Lasota, On traveling round without feeling it and uncurving curves, *American Journal of Physics* **54**, pp. 936–939, 1986. Cited on page 450.

425 About black holes in the early universe, see ... Cited on page 451.

426 About black holes formation via star collapse, see ... Cited on page 451.

427 Frederick Lamb, APS meeting 1998, binary star 4U1820-30, 20 000 lightyears from Earth, Physics News Update, April 27, 1998. Cited on page 452.

428 The first evidence for the observation of matter falling into a black hole was publicised in early 2001. ... Cited on page 452.

429 For a readable summary of the Penrose–Hawking singularity theorems, see ... Details can be found in reference Ref. 392. Cited on page 452.

430 For an overview of cosmic censorship, see T.P. Singh, Gravitational collapse, black holes and naked singularities, http://www.arxiv.org/abs/gr-qc/9805066 or R.M. Wald, Gravitational collapse and cosmic censorship, http://www.arxiv.org/abs/gr-qc/9710068. The original idea is due to R. Penrose, Rivista del Nuovo Cimento, *1* p. 252, 1969. Cited on page 453.

431 The geometrodynamic clock is discussed in D.E. Brahm & R.P. Gruber, Limitations of the geometrodynamic clock, *General Relativity and Gravitation* **24**, pp. 297–303, 1992. The clock itself was introduced by R.F. Marzke, in his Ph.D. thesis The theory of measurement in general relativity, 1959, performed with John Wheeler as thesis adviser. Cited on page 456.

432 G.J. Stoney, On the Physical Units of Nature, *Philosophical Magazine* **11**, pp. 381–391, 1881. Cited on page 455.

433 R. Geroch, Einstein Algebras, *Commun. Math. Phys.* **26**, pp. 271–275, 1972. Cited on page 456.

434 A. Macdonald, Einstein's hole argument, *American Journal of Physics* **69**, pp. 223–225, 2001. Cited on page 457.

435 Roman U. Sexl, Die Hohlwelttheorie, *Der mathematisch-naturwissenschaftliche Unterricht* **368**, pp. 453–460, 1983. Roman U. Sexl, Universal Conventionalism and Space-Time., *General Relativity and Gravitation* **1**, pp. 159–180, 1970. See also Roman U. Sexl, Die Hohlwelttheorie, pp. 241–258, in Arthur Scharmann & Herbert Schramm, editors, *Physik, Theorie, Experiment, Geschichte, Didaktik – Festschrift für Wilfried Kuhn zum 60. Geburtstag am 6. Mai 1983*, Aulis Verlag Deubner 1984. Cited on page 458.

436 T. Damour, Experimental tests of relativistic gravity, http://www.arxiv.org/abs/gr-qc/9904057. It is the latest in a series of his papers on the topic; the first was T. Damour, Was Einstein 100 % right?, http://www.arxiv.org/abs/gr-qc/9412064. Cited on pages 461 and 462.

437 H. Dittus, F. Everitt, C. Lämmerzahl & G. Schäfer, Die Gravitation im Test, *Physikalische Blätter* **55**, pp. 39–46, 1999. Cited on page 461.

438 The inverse square dependence has been checked down to 60 μm, as reported by E. Adelberger, B. Heckel & C.D. Hoyle, Testing the gravitational inverse-square law, *Physics World* 2005. Cited on page 462.

439 For theories competing with general relativity, see for example C.M. Will, The Confrontation between General Relativity and Experiment, *Living Reviews of Relativity* 2001, http://www.livingreviews.org/lrr-2001-4. Among others, the absence of the Nordtvedt effect, a hypothetical 28-day oscillation in the Earth–Moon distance, looked for by laser ranging experiments without any result, 'killed' several competing theories. This effect, predicted by Kenneth Nordtvedt, would only appear if the gravitational energy in the Earth–Moon system would fall differently from the Earth and the Moon themselves. For a summary of the measurements, see J. Müller, M. Schneider, M. Soffel & H. Ruder, *Astrophysical Journal Letters* 382, p. L101, 1991. Cited on page 461.

440 See S. Bässler & al., Improved test of the equivalence principle for gravitational selfenergy, *Physical Review Letters* 83, pp. 3585–3588, 1st November 1999. See also C.M. Will, Gravitational radiation and the validity of general relativity, *Physics Today* 52, p. 38, October 1999. No citations.

441 Almost everything of importance in general relativity is published in the Journal *Classical and quantum gravity*. For the US American side of research, see the webzine Matters of Gravity, published on http://vishnu.nirvana.phys.psu.edu/mog.html. Cited on page 461.

442 Collisions and many body problems ... Cited on page 462.

443 Inflation and early universe ... Cited on page 462.

444 The study of chaos in Einstein's field equations is in its beginnings. See e.g. L. Bombelli, F. Lombardo & M. Castagnino, Chaos in Robertson-Walker Cosmology, electronic preprint available at http://www.arxiv.org/abs/gr-qc/9707051. Cited on page 462.

445 The ESA satellite called 'Planck' will measure the polarization of the cosmic microwave background. Cited on page 462.

446 A good introduction into the topic of gamma ray bursts is S. Klose, J. Greiner & D. Hartmann, Kosmische Gammastrahlenausbrüche – Beobachtungen und Modelle, Teil I und II, *Sterne und Weltraum* March and April 2001. Cited on page 462.

447 The field solution database is built around the work of A. Karlhede, which allows one to distinguish solutions with a limited amount of mathematical computation. Cited on page 462.

448 Torsion is presented in ... Cited on page 463.

449 Wormholes and nontrivial topologies ... A basic approach is the one by Mark Hadley, *Classical and Quantum Gravity* October 2000. Cited on page 463.

450 Other formulations of relativity ... Cited on page 463.

451 Teaching will benefit in particular from new formulations, from concentration on principles and their consequences, as has happened in special relativity, from simpler descriptions at the weak field level, and from future research in the theory of general relativity. The newer textbooks cited above all are steps in these directions. Cited on page 463.

452 A beautiful introduction into modern astronomy is Paolo Maffei, *I mostri del cielo*, Mondadori Editore, 1976. No citations.

453 Michael H. Soffel, *Relativity in Astronomy, Celestial Mechanics and Geodesy*, Springer Verlag, 1989. No citations.

454 M. Nagano & A.A. Watson, Observations and implications of the ultrahigh energy cosmic rays, *Reviews of Modern Physics* 72, pp. 689–732, 2000. No citations.

455 The speed of light as function of height has been checked e.g. by ... The change is smaller than ... No citations.

456 Steven Weinberg, The cosmological constant problem, *Reviews of Modern Physics* **61**, pp. 1–23, 1989. This influential review article on the possible views and solutions is frequently cited, though now outdated. No citations.

457 A. Ashtekar, 1986, New variables for classical and quantum gravity, *Physical Review Letters* 57, pp. 2244–2247, No citations.

458 A well written text on the connections between the big bang and particle physics is by I.L. Rozental, *Big Bang – Big Bounce, How Particles and Fields Drive Cosmic Evolution*, Springer, 1988. No citations.

Chapter IV

Classical Electrodynamics

What is light? The study of relativity left us completely in the dark, even though we had embarked in it precisely to find an answer to that question. True, we have learned how the motion of light compares with that of objects. We also learned that light is a moving entity that cannot be stopped; but we haven't learned anything about its nature. The answer to this long-standing question emerges only from the study of those types of motion that are *not* related to gravitation, such as the ways magicians levitate objects.

14. Liquid electricity, invisible fields and maximum speed

Page 137 Revisiting the list of motors found in this world, we remark that gravitation hardly describes any of them. Neither the motion of sea waves, fire and earthquakes, nor that of a gentle breeze are due to gravity. The same applies to the motion of muscles. Have you ever listened to your own heart beat with a stethoscope? Without having done so, you cannot claim to have experienced the mystery of motion. Your heart has about 3000 million beats in your lifetime. Then it stops. Challenge 885 e

It was one of the most astonishing discoveries of science that heart beats, sea waves and most other cases of everyday motion, as well as the nature of light itself, are connected to observations made thousands of years ago using two strange stones. These stones show that all examples of motion, which are called *mechanical* in everyday life, are, without exception, of *electrical* origin.

In particular, the solidity, the softness and the impenetrability of matter are due to internal electricity; also the emission of light is an electrical process. As these aspects are part of everyday life, we will leave aside all complications due to gravity and curved space-time. The most productive way to study electrical motion is to start, as in the case of gravity, with those types of motion which are generated without any contact between the bodies involved.

Amber, lodestone and mobile phones

> Any fool can ask more questions than seven sages can answer.

The story of electricity starts with trees. Trees have a special relation to electricity. When a tree is cut, a viscous resin appears. With time it solidifies and, after millions of years, it forms *amber*. When amber is rubbed with a cat fur, it acquires the ability to attract small objects, such as saw dust or pieces of paper. This was already known to Thales of Miletus,

Figure 214 Objects surrounded by fields: amber, lodestone and mobile phone

one of the original seven sages, in the sixth century BCE. The same observation can be made with many other polymer combinations, for example with combs and hair, with soles of the shoe on carpets, and with a TV screen and dust. Children are always surprised by the effect, shown in Figure 215, that a rubbed comb has on running tap water. Another interesting effect can be observed when a rubbed comb is put near a burning candle. (Can you imagine what happens?)

Challenge 886 ny

Another part of the story of electricity involves an iron mineral found in certain caves around the world, e.g. in a region (still) called Magnesia in the Greek province of Thessalia, and in some regions in central Asia. When two stones of this mineral are put near each other, they attract or repel each other, depending on their relative orientation. In addition, these stones attract objects made of cobalt, nickel or iron.

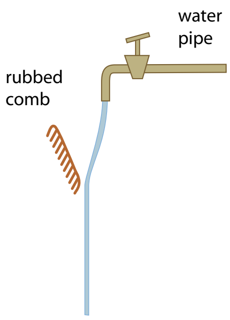


Figure 215 How to amaze kids

Today we also find various small objects in nature with more sophisticated properties, as shown in Figure 214. Some objects enable you to switch on a television, others unlock car doors, still others allow you to talk with far away friends.

All these observations show that in nature there are situations where bodies exert influence on others *at a distance*. The space surrounding a body exerting such an influence is said to contain a field. A *(physical) field* is thus an entity that manifests itself by accelerating other bodies in its region of space. A field is some 'stuff' taking up space. Experiments show that fields have no mass. The field surrounding the mineral found in Magnesia is called a *magnetic field* and the stones are called *magnets*.* The field around amber – called ἤλεκτρον in Greek, from a root meaning 'brilliant, shining' – is called an *electric field*. The name is due to a proposal by the famous English part-time physicist William Gilbert (1544–1603) who was physician to Queen Elizabeth I. Objects surrounded by a permanent electric field are called *electrets*. They are much less common than magnets; among others, they are used in certain loudspeaker systems.**

* A pretty book about the history of magnetism and the excitement it generates is JAMES D. LIVINGSTON, *Driving Force – the Natural Magic of Magnets*, Harvard University Press, 1996.

** The Kirlian effect, which allows one to make such intriguingly beautiful photographs, is due to a time-

Table 39 Searches for magnetic monopoles, i.e., for magnetic charges

Search	Magnetic charge
Smallest magnetic charge suggested by quantum theory	$g = \frac{h}{e} = \frac{eZ_0}{2\alpha} = 4.1\,\mathrm{pWb}$
Search in minerals	none Ref. 460
Search in meteorites	none Ref. 460
Search in cosmic rays	none Ref. 460
Search with particle accelerators	none Ref. 460

The field around a mobile phone is called a *radio* field or, as we will see later, an *electromagnetic* field. In contrast to the previous fields, it oscillates over time. We will find out later that many other objects are surrounded by such fields, though these are often very weak. Objects that emit oscillating fields, such as mobile phones, are called radio transmitters or radio emitters.

Fields influence bodies over a distance, without any material support. For a long time, this was rarely found in everyday life, as most countries have laws to restrict machines that use and produce such fields. The laws require that for any device that moves, produces sound, or creates moving pictures, the fields need to remain inside them. For this reason a magician moving an object on a table via a hidden magnet still surprises and entertains his audience. To feel the fascination of fields more strongly, a deeper look into a few experimental results is worthwhile.

How can one make lightning?

Everybody has seen a lightning flash or has observed the effect it can have on striking a tree. Obviously lightning is a moving phenomenon. Photographs such as that of Figure 216 show that the tip of a lightning flash advance with an average speed of around 600 km/s. But *what* is moving? To find out, we have to find a way of making lightning for ourselves.

In 1995, the car company General Motors accidentally rediscovered an old and simple method of achieving this. Their engineers had inadvertently built a spark generating mechanism into their cars; when filling the petrol tank, sparks were generated, which sometimes lead to the explosion of the fuel. They had to recall 2 million vehicles of its Opel brand. Ref. 461

What had the engineers done wrong? They had unwittingly copied the conditions for a electrical device which anyone can build at home and which was originally invented by William Thomson.* Repeating his experiment today, we would take two water taps, four

varying electric field.

* William Thomson (1824–1907), important Irish Unionist physicist and professor at Glasgow University. He worked on the determination of the age of the Earth, showing that it was much older than 6000 years, as several sects believed. He strongly influenced the development of the theory of magnetism and electricity, the description of the aether and thermodynamics. He propagated the use of the term 'energy' as it is used today, instead of the confusing older terms. He was one of the last scientists to propagate mechanical analogies for the explanation of phenomena, and thus strongly opposed Maxwell's description of electromagnetism. It

Table 40 Some observed magnetic fields

Observation	Magnetic field
Lowest measured magnetic field (e.g., fields of the Schumann resonances)	1 fT
Magnetic field produced by brain currents	0.1 pT to 3 pT
Intergalactic magnetic fields	1 pT to 10 pT
Magnetic field in the human chest, due to heart currents	100 pT
Magnetic field of our galaxy	0.5 nT
Magnetic field due to solar wind	0.2 to 80 nT
Magnetic field directly below high voltage power line	0.1 to 1 µT
Magnetic field of Earth	20 to 70 µT
Magnetic field inside home with electricity	0.1 to 100 µT
Magnetic field near mobile phone	100 µT
Magnetic field that influences visual image quality in the dark	100 µT
Magnetic field near iron magnet	100 mT
Solar spots	1 T
Magnetic fields near high technology permanent magnet	max 1.3 T
Magnetic fields that produces sense of coldness in humans	5 T or more
Magnetic fields in particle accelerator	10 T
Maximum static magnetic field produced with superconducting coils	22 T
Highest static magnetic fields produced in laboratory, using hybrid magnets	45 T
Highest *pulsed* magnetic fields produced without coil destruction	76 T
Pulsed magnetic fields produced, lasting about 1 µs, using imploding coils	1000 T
Field of white dwarf	10^4 T
Fields in petawatt laser pulses	30 kT
Field of neutron star	from 10^6 T to 10^{11} T
Quantum critical magnetic field	4.4 GT
Highest field ever measured, on magnetar and soft gamma repeater SGR-1806-20	0.8 to $1 \cdot 10^{11}$ T
Field near nucleus	1 TT
Maximum (Planck) magnetic field	$2.2 \cdot 10^{53}$ T

empty bean or coffee cans, of which two have been opened at both sides, some nylon rope and some metal wire. Ref. 462

Putting this all together as shown in Figure 217, and letting the water flow, we find a strange effect: large sparks periodically jump between the two copper wires at the point

was mainly for this reason that he failed to receive a Nobel prize. He was also one of the minds behind the laying of the first transatlantic telegraphic cable. Victorian to his bones, when he was knighted, he chose the name of a small brook near his home as his new name; thus he became Lord Kelvin of Largs. Therefore the unit of temperature obtained its name from a small Scottish river.

Figure 216 Lightning: a picture taken with a moving camera, showing its multiple strokes (© Steven Horsburgh)

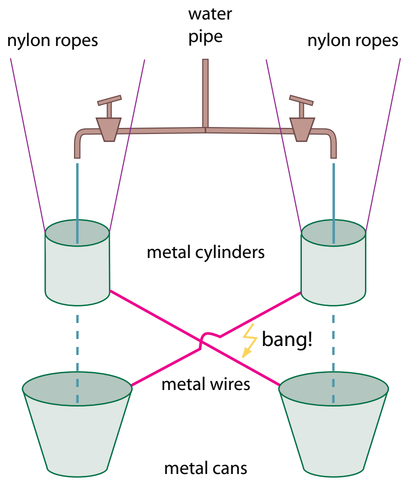


Figure 217 A simple Kelvin generator

where they are nearest to each other, giving out loud bangs. Can you guess what condition for the flow has to be realized for this to work? And what did Opel do to repair the cars they recalled?

Challenge 887 n

If we stop the water flowing just before the next spark is due, we find that both buckets are able to attract sawdust and pieces of paper. The generator thus does the same that rubbing amber does, just with more bang for the buck(et). Both buckets are surrounded by electric fields. The fields increase with time, until the spark jumps. Just after the spark, the buckets are (almost) without electric field. Obviously, the flow of water somehow builds up an entity on each bucket; today we call this *electric charge*. Charge can flow in metals and, when the fields are high enough, through air. We also find that the two buckets are surrounded by two different types of electric fields: bodies that are attracted by one bucket are repelled by the other. All other experiments confirm that there are *two* types of charges. The US politician and part-time physicist Benjamin Franklin (1706–1790) called the electricity created on a glass rod rubbed with a dry cloth *positive*, and that on a piece of amber *negative*. (Previously, the two types of charges were called 'vitreous' and 'resinous'.) Bodies with charges of the same sign repel each other, bodies with opposite charges attract each other; charges of opposite sign flowing together cancel each other out.*

In summary, electric fields start at bodies, provided they are charged. Charging can be achieved by rubbing and similar processes. Charge can flow: it is then called an electric *current*. The worst conductors of current are polymers; they are called insulators or dielectrics. A charge put on an insulator remains at the place where it was put. In contrast, metals are good conductors; a charge placed on a conductor spreads all over its surface. The best conductors are silver and copper. This is the reason that at present, after a hundred years of use of electricity, the highest concentration of copper in the world is below the surface of Manhattan.

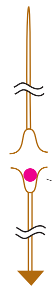


Figure 218 Franklin's personal lightning rod

Of course, one has to check whether natural lightning is actually electrical in origin. In 1752, experiments performed in France, following a suggestion by Benjamin Franklin, published in London in 1751, showed that one can indeed draw electricity from a thunderstorm via a long rod.* These French experiments made Franklin famous worldwide; they were also the start of the use of lightning rods all over the world. Later, Franklin had a lightning rod built through his own house, but of a somewhat unusual type, as shown in Figure 218. Can you guess what it did in his hall during bad weather, all parts being made of metal? (Do not repeat this experiment; the device can kill.)

Challenge 888 n

Electric charge and electric fields

If all experiments with charge can be explained by calling the two charges positive and negative, the implication is that some bodies have more, and some less charge than an uncharged, *neutral* body. Electricity thus only flows when two differently charged bodies are brought into contact. Now, if charge can flow and accumulate, we must be able to somehow measure its amount. Obviously, the *amount* of charge on a body, usually abbreviated q, is defined via the influence the body, say a piece of sawdust, feels when subjected to a

* In fact, there are many other ways to produces sparks or even *arcs*, i.e. sustained sparks; there is even a complete subculture of people who do this as a hobby at home. Those who have a larger budget do it professionally, in particle accelerators. See the http://www.kronjaeger.com/hv/ website.

* The details of how lightning is generated and how it propagates are still a topic of research. An introduction is given on page 551.

Table 41 Properties of *classical* electric charge

Electric charges	Physical property	Mathematical name	Definition
Can be distinguished	distinguishability	element of set	Page 599
Can be ordered	sequence	order	Page 1099
Can be compared	measurability	metricity	Page 1108
Can change gradually	continuity	completeness	Page 1116
Can be added	accumulability	additivity	Page 66
Do not change	conservation	invariance	$q = \text{const}$
Can be separated	separability	positive or negative	

field. Charge is thus defined by comparing it to a standard reference charge. For a charged body of mass m accelerated in a field, its charge q is determined by the relation

$$\frac{q}{q_{\text{ref}}} = \frac{ma}{m_{\text{ref}} a_{\text{ref}}} \, , \tag{388}$$

i.e., by comparing it with the acceleration and mass of the reference charge. This definition reflects the observation that mass alone is not sufficient for a complete characterization of a body. For a full description of motion we need to know its electric charge; charge is therefore the second intrinsic property of bodies that we discover in our walk.

Nowadays the unit of charge, the *coulomb*, is defined through a standard flow through metal wires, as explained in Appendix B. This is possible because all experiments show that charge is conserved, that it flows, that it flows continuously and that it can accumulate. Charge thus behaves like a fluid substance. Therefore we are forced to use for its description a scalar quantity q, which can take positive, vanishing, or negative values.

In everyday life these properties of electric charge, listed also in Table 41, describe observations with sufficient accuracy. However, as in the case of all previously encountered classical concepts, these experimental results for electrical charge will turn out to be only approximate. More precise experiments will require a revision of several properties. However, no counterexample to charge conservation has as yet been observed.

A charged object brought near a neutral one polarizes it. *Electrical polarization* is the separation of the positive and negative charges in a body. For this reason, even neutral objects, such as hair, can be attracted to a charged body, such as a comb. Generally, both insulators and conductors can be polarized; this occurs for whole stars down to single molecules.

Attraction is a form of acceleration. Experiments show that the entity that accelerates charged bodies, the *electric field*, behaves like a small arrow fixed at each point $\mathbf{x}$ in space; its length and direction do not depend on the observer. In short, the electric field $\mathbf{E}(\mathbf{x})$ is a *vector* field. Experiments show that it is best defined by the relation

$$q\mathbf{E}(\mathbf{x}) = m\mathbf{a}(\mathbf{x}) \tag{389}$$

Table 42 Values of electrical charge observed in nature

Observation	Charge
Smallest measured non-vanishing charge	$1.6 \cdot 10^{-19}$ C
Charge per bit in computer memory	10^{-13} C
Charge in small capacitor	10^{-7} C
Charge flow in average lightning stroke	1 C to 100 C
Charge stored in a fully-charge car battery	0.2 MC
Charge of planet Earth	1 MC
Charge separated by modern power station in one year	$3 \cdot 10^{11}$ C
Total charge of positive (or negative) sign observed in universe	$10^{62\pm2}$ C
Total charge observed in universe	0 C

Table 43 Some observed electric fields

Observation	Electric field
Field 1 m away from an electron in vacuum	Challenge 891 n
Field values sensed by sharks	down to 0.1 μV/m
Cosmic noise	10 μV/m
Field of a 100 W FM radio transmitter at 100 km distance	0.5 mV/m
Field inside conductors, such as copper wire	0.1 V/m
Field just beneath a high power line	0.1 to 1 V/m
Field of a GSM antenna at 90 m	0.5 V/m
Field inside a typical home	1 to 10 V/m
Field of a 100 W bulb at 1 m distance	50 V/m
Ground field in Earth's atmosphere	100 to 300 V/m
Field inside thunder clouds	up to over 100 kV/m
Maximum electric field in air before sparks appear	1 to 3 MV/m
Electric fields in biological membranes	10 MV/m
Electric fields inside capacitors	up to 1 GV/m
Electric fields in petawatt laser pulses	10 TV/m
Electric fields in U^{91+} ions, at nucleus	1 EV/m
Maximum practical electric field in vacuum, limited by electron pair production	1.3 EV/m
Maximum possible electric field in nature (corrected Planck electric field)	$2.4 \cdot 10^{61}$ V/m

Challenge 890 e

taken at every point in space **x**. The definition of the electric field is thus based on how it *moves* charges.* The field is measured in multiples of the unit N/C or V/m.

Challenge 889 ny

* Does the definition of electric field given here assume a charge speed that is much less than that of light?

To describe the motion due to electricity completely, we need a relation explaining how charges *produce* electric fields. This relation was established with precision (but not for the first time) by Charles-Augustin de Coulomb on his private estate, during the French Revolution.* He found that around any small-sized or any spherical charge Q *at rest* there is an electric field. At a position $\mathbf{r}$, the electric field $\mathbf{E}$ is given by

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\frac{\mathbf{r}}{r} \quad \text{where} \quad \frac{1}{4\pi\varepsilon_0} = 9.0\,\mathrm{GV\,m/C}\,. \tag{390}$$

Later we will extend the relation for a charge in motion. The bizarre proportionality constant, built around the so-called *permittivity of free space* ε_0, is due to the historical way the unit of charge was defined first.* The essential point of the formula is the decrease of

Challenge 892 n the field with the square of the distance; can you imagine the origin of this dependence?

The two previous equations allow one to write the interaction between two charged bodies as

$$\frac{\mathrm{d}\mathbf{p}_1}{\mathrm{d}t} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}\frac{\mathbf{r}}{r} = -\frac{\mathrm{d}\mathbf{p}_2}{\mathrm{d}t}\,, \tag{391}$$

where $\mathrm{d}\mathbf{p}$ is the momentum change, and $\mathbf{r}$ is the vector connecting the two centres of mass. This famous expression for electrostatic attraction and repulsion, also due to Coulomb, is valid only for charged bodies that are of small size or spherical, and most of all, that are *at rest*.

Electric fields have two main properties: they contain energy and they can polarize bodies. The energy content is due to the electrostatic interaction between charges. The strength of the interaction is considerable. For example, it is the basis for the force of our muscles. Muscular force is a macroscopic effect of equation 391. Another example is the material strength of steel or diamond. As we will discover, all atoms are held together by electrostatic attraction. To convince yourself of the strength of electrostatic attraction, answer the following: What is the force between two boxes with a gram of protons each,

Challenge 893 n located on the two poles of the Earth? Try to guess the result before you calculate the astonishing value.

Coulomb's relation for the field around a charge can be rephrased in a way that helps to generalize it to non-spherical bodies. Take a closed surface A, i.e., a surface than encloses a certain volume. Then the integral of the electric field over this surface is the enclosed charge Q divided by ε_0:

$$\int_{\text{closed surface}} \boldsymbol{E}\,\mathrm{d}\boldsymbol{A} = \frac{Q}{\varepsilon_0}\,. \tag{392}$$

This mathematical relation follows from the result of Coulomb and is called *Gauss's law*. It is strictly valid only for static situations. Since inside conductors the electrical field is

* Charles-Augustin de Coulomb (b. 1736 Angoulême, d. 1806 Paris), French engineer and physicist. His careful experiments on electric charges provided a firm basis for the study of electricity.

* Other definitions of this and other proportionality constants to be encountered later are possible, leading to *unit systems* different from the SI system used here. The SI system is presented in detail in Appendix B. Among the older competitors, the Gaussian unit system often used in theoretical calculations, the Heaviside–Lorentz unit system, the electrostatic unit system and the electromagnetic unit system are

Ref. 463 the most important ones.

zero, the law implies, for example, that if a charge q is surrounded by a metal sphere, the
Challenge 894 e outer surface of the metal sphere shows the same charge q.

Owing to the strength of electromagnetic interactions, separating charges is not an easy task. This is the reason that electrical effects have only been commonly used for about a hundred years. We had to wait for practical and efficient devices to be invented for separating charges and putting them into motion. Of course this implies that energy is needed. Batteries, as used in mobile phones, use chemical energy to do the trick.* Thermoelectric elements, as used in some watches, use the temperature difference between the wrist and the air to separate charges; solar cells use light, and dynamos or Kelvin generators use kinetic energy.

Do uncharged bodies attract one other? In first approximation they do not. But when the question is investigated more precisely, one finds that they can attract one other. Can
Challenge 896 n you find the conditions for this to happen? In fact, the conditions are quite important, as our own bodies, which are made of neutral molecules, are held together in this way.

What then is electricity? The answer is simple: *electricity is nothing in particular*. It is the name for a field of inquiry, but not the name for any specific observation or effect. Electricity is neither electric current, nor electric charge, nor electric field. Electricity is not a specific term; it applies to *all* of these phenomena. In fact the vocabulary issue hides a deeper question that remains unanswered at the beginning of the twenty-first century:what is the nature of electric charge? In order to reach this issue, we start with the following question.

Can we detect the inertia of electricity?

If electric charge really is something *flowing* through metals, we should be able to observe the effects shown in Figure 219. Maxwell has predicted most of these effects: electric charge should fall, have inertia and be separable from matter. Indeed, each of these effects has been observed.** For example, when a long metal rod is kept vertically, we can measure an electrical potential difference, a voltage, between the top and the bottom. In other words, we can measure the *weight* of electricity in this way. Similarly, we can meas-
Ref. 464 ure the potential difference between the ends of an accelerated rod. Alternatively, we can measure the potential difference between the centre and the rim of a rotating metal disc. The last experiment was, in fact, the way in which the ratio q/m for currents in metals was first measured with precision. The result is

$$q/m = 1.8 \cdot 10^{11}\ \mathrm{C/kg} \qquad (393)$$

for all metals, with small variations in the second digit. In short, electrical current has mass. Therefore, whenever we switch on an electrical current, we get a *recoil*. This simple
Ref. 465 effect can easily be measured and confirms the mass to charge ratio just given. Also, the emission of current into air or into vacuum is observed; in fact, every television tube uses this principle to generate the beam producing the picture. It works best for metal
Ref. 466 objects with sharp, pointed tips. The rays created this way – we could say that they are

Challenge 895 n * Incidentally, are batteries sources of charges?

** Maxwell also performed experiments to detect these effects (apart from the last one, which he did not predict), but his apparatuses where not sensitive enough.

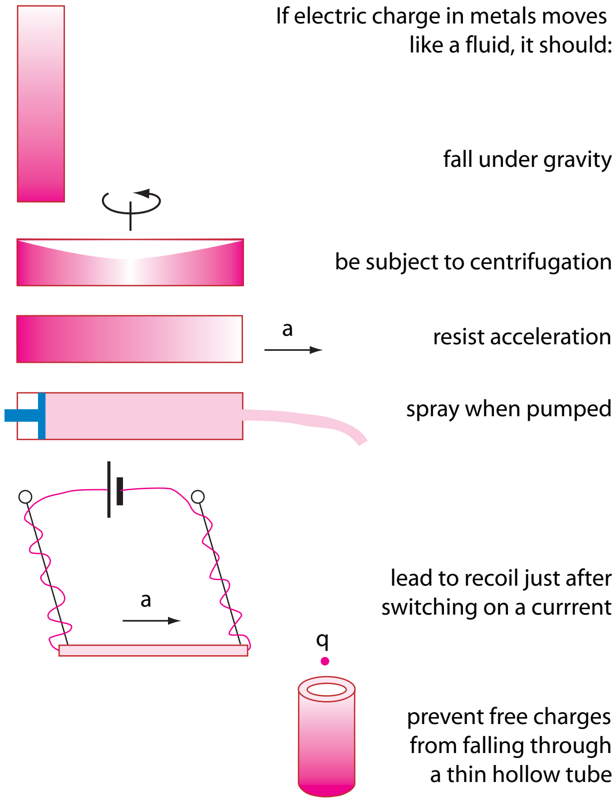


Figure 219 Consequences of the flow of electricity

'free' electricity – are called *cathode rays*. Within a few per cent, they show the same mass to charge ratio as expression (393). This correspondence thus shows that charges move almost as freely in metals as in air; this is the reason metals are such good conductors.

If electric charge *falls* inside vertical metal rods, we can make the astonishing deduction that cathode rays – as we will see later, they consist of free electrons* – should not be able to fall through a vertical metal tube. This is due to exact compensation of the acceleration by the electrical field generated by the displaced electricity in the tube and Challenge 897 e the acceleration of gravity. Thus electrons should not be able to fall through a long thin cylinder. This would not be the case if electricity in metals did not behave like a fluid. The Ref. 467 experiment has indeed been performed, and a reduction of the acceleration of free fall for electrons of 90 % has been observed. Can you imagine why the ideal value of 100 % is Challenge 898 n not achieved?

* The name 'electron' is due to George Stoney. Electrons are the smallest and lightest charges moving in metals; they are, usually – but not always – the 'atoms' of electricity – for example in metals. Their charge is small, 0.16 aC, so that flows of charge typical of everyday life consist of large numbers of electrons; as a result, electrical charge behaves like a continuous fluid. The particle itself was discovered and presented in 1897 by the Prussian physicist Johann Emil Wiechert (1861–1928) and, independently, three months later, by the British physicist Joseph John Thomson (1856–1940).

Table 44 Some observed electric current values

Observation	Current
Smallest regularly measured currents	1 fA
Human nerve signals	20 μA
Lethal current for humans	as low as 20 mA, typically 100 mA
Current drawn by a train engine	600 A
Current in a lightning bolt	10 to 100 kA
Highest current produced by humans	20 MA
Current inside the Earth	around 100 MA

Feeling electric fields

Why is electricity dangerous to humans? The main reason is that the human body is controlled by 'electric wires' itself. As a result, outside electricity interferes with the internal signals. This has been known since 1789. In that year the Italian medical doctor Luigi Galvani (1737–1798) discovered that electrical current makes the muscles of a dead animal contract. The famous first experiment used frog legs: when electricity was applied to them, they twitched violently. Subsequent investigations confirmed that all nerves make use of electrical signals. Nerves are the 'control wires' of animals. However, nerves are not made of metal: metals are not sufficiently flexible. As a result, nerves do not conduct electricity using electrons but by using ions. The finer details were clarified only in the twentieth century. Nerve signals propagate using the motion of sodium and potassium ions in the cell membrane of the nerve. The resulting signal speed is between 0.5 m/s and 120 m/s, depending on the type of nerve. This speed is sufficient for the survival of most species – it signals the body to run away in case of danger.

Being electrically controlled, all mammals can sense strong electric fields. Humans can sense fields down to around 10 kV/m, when hair stands on end. In contrast, several animals can sense weak electric and magnetic fields. Sharks, for example, can detect fields down to 1 μV/m using special sensors, the Ampullae of Lorenzini, which are found around their mouth. Sharks use them to detect the field created by prey moving in water; this allows them to catch their prey even in the dark. Several freshwater fish are also able to detect electric fields. The salamander and the platypus, the famous duck-billed mammal, can also sense electric fields. Like sharks, they use them to detect prey in water which is too muddy to see through. Certain fish, the so-called *weakly-electric fish*, even generate a weak field in order to achieve better prey detection.*

No land animal has special sensors for electric fields, because any electric field in air is strongly damped when it encounters a water-filled animal body. Indeed, the usual atmosphere has an electric field of around 100 V/m; inside the human body this field is damped to the μV/m range, which is much less than an animal's internal electric fields.

* It took until the year 2000 for technology to make use of the same effect. Nowadays, airbag sensors in cars often use electric fields to sense whether the person sitting in the seat is a child or an adult, thus changing the way that the bag behaves in an accident.

In other words, humans do not have sensors for low electric fields because they are land animals. (Do humans have the ability to sense electric fields in water? Nobody seems to know.) However, there a few exceptions. You might know that some older people can sense approaching thunderstorms in their joints. This is due the coincidence between the electromagnetic field frequency emitted by thunderclouds – around 100 kHz – and the resonant frequency of nerve cell membranes.

Challenge 899 ny
Page 521

The water content of the human body also means that the electric fields in air that are found in nature are rarely dangerous to humans. Whenever humans do sense electric fields, such as when high voltage makes their hair stand on end, the situation is potentially dangerous.

The high impedance of air also means that, in the case of time-varying electromagnetic fields, humans are much more prone to be affected by the magnetic component than by the electric component.

Magnets

The study of magnetism progressed across the world independently of the study of electricity. Towards the end of the 12th century, the compass came into use in Europe. At that time, there were heated debates on whether it pointed to the north or the south. In 1269, the French military engineer Pierre de Maricourt (1219–1292) published his study of magnetic materials. He found that every magnet has *two* points of highest magnetization, and he called them *poles*. He found that even after a magnet is cut, the resulting pieces always retain two poles: one points to the north and the other to the south when the stone is left free to rotate. Magnets are dipoles. There are no magnetic monopoles. Despite the promise of eternal fame, no magnetic monopole has ever been found, as shown in Table 39.

Ref. 468

Magnets have a second property: magnets transform unmagnetic objects into magnetic ones. There is thus also a magnetic polarization, similar to the electric polarization.

Can humans feel magnetic fields?

It is known that honey bees, sharks, pigeons, salmon, trout, sea turtles and certain bacteria can feel magnetic fields. One speaks of the ability for magnetoreception. All these life forms use this ability for navigation. The most common detection method is the use of small magnetic particles inside a cell; the cell then senses how these small built-in magnets move in a magnetic field. The magnets are tiny, typically around 50 nm in size. These small magnets are used to navigate along the magnetic field of the Earth. For higher animals, the variations of the magnetic field of the Earth, 20 to 70 μT, produce a landscape that is similar to the visible landscape for humans. They can remember it and use it for navigation.

Ref. 476

Can humans feel magnetic fields? Magnetic material seems to be present in the human brain, but whether humans can feel magnetic fields is still an open issue. Maybe you can devise a way to check this?

Challenge 900 ny

Are magnetism and electricity related? François Arago* found out that they were. He observed that a ship that had survived a bad thunderstorm and had been struck by light-

* Dominique-François Arago (1786–1853) French physicist.

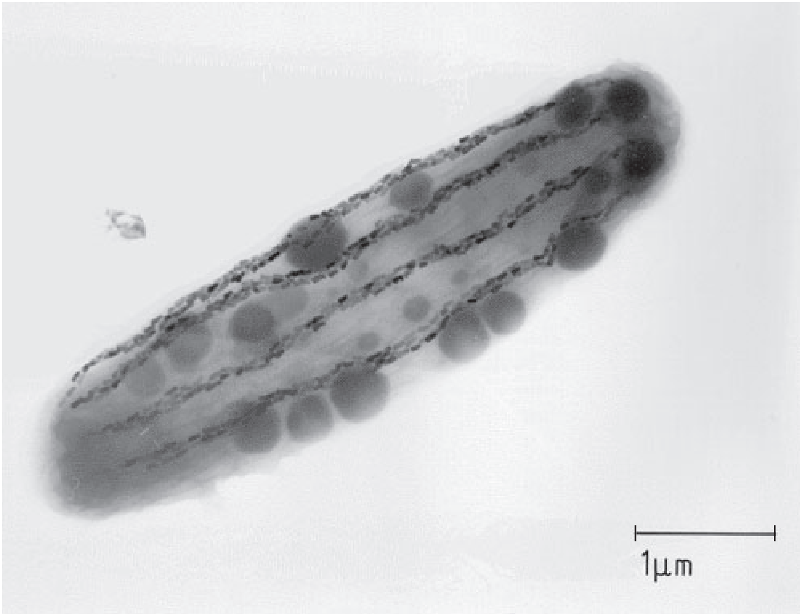


Figure 220 The magentotactic bacterium *Magnetobacterium bavaricum* with its magnetosomes (photograph by Marianne Hanzlik)

ning, needed a new compass. Thus lightning has the ability to demagnetise compasses. Arago knew, like Franklin, that lightning is an electrical phenomena. In other words, electricity and magnetism must be related. More precisely, magnetism must be related to the *motion* of electricity.

How can one make a motor?

Communism is soviets plus electricity.

Lenin*

The reason for Lenin's famous statement were two discoveries. One was made in 1820 by the Danish physicist Hans Christian Oersted (1777–1851) and the other in 1831 by the English physicist Michael Faraday.** The consequences of these experiments changed the world completely in less than one century.

On the 21st of July of 1821, Oersted published a leaflet, in Latin, which took Europe by storm. Oersted had found (during a lecture demonstration to his students) that when a

* Lenin (b. 1870 Simbirsk, d. 1924 Gorki), founder of the Soiet Union.

** Michael Faraday (b. 1791 Newington Butts, d. 1867 London) was born to a simple family, without schooling, and of deep and naive religious ideas. As a boy he became assistant to the most famous chemist of his time, Humphry Davy (1778–1829). He had no mathematical training, but late in his life he became member of the Royal Society. A modest man, he refused all other honours in his life. He worked on chemical topics, the atomic structure of matter and, most of all, developed the idea of (magnetic) fields and field lines through all his experimental discoveries, such as effect. Fields were later described mathematically by Maxwell, who at that time was the only person in Europe to take over Faraday's field concept.

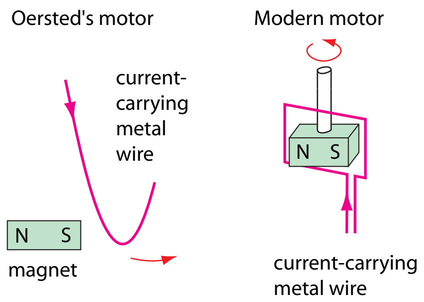


Figure 221 An old and a newer version of an electric motor

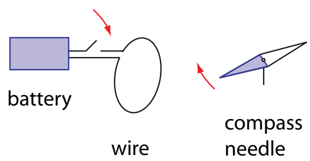


Figure 222 An electrical current always produces a magnetic field

current is sent through a wire, a nearby magnet is put into motion. In other words, he found that the flow of electricity can move bodies.

Further experiments show that *two* wires in which charges flow attract or repel each other, depending on whether the currents are parallel or antiparallel. These and other experiments show that wires in which electricity flows behave like magnets.* In other words, Oersted had found the definite proof that electricity could be turned into magnetism.

Shortly afterwards, Ampère* found that *coils* increase these effects dramatically. Coils behave like small magnets. In particular, coils, like magnetic fields, always have two poles, usually called the north and the south pole. Opposite poles attract, like poles repel each other. As is well known, the Earth is itself a large magnet, with its magnetic north pole near the geographic south pole, and vice versa.

Moving electric charge produces magnetic fields. This result explains why magnetic fields always have two poles. The lack of magnetic monopoles thus becomes clear. But one topic is strange. If magnetic fields are due to the motion of charges, this must be also the case for a normal magnet. Can this be shown?

In 1915, two men in the Netherlands found a simple way to prove that even in a magnet, something is moving. They suspended a metal rod from the ceiling by a thin thread and then put a coil around the rod, as shown in Figure 223. They predicted that the tiny currents inside the rod would become aligned by the magnetic field of the coil. As a result, they expected that a current passing through the coil would make the rod turn around its axis. Indeed, when they sent a strong current through the coil, the rod rotated. (As a result of the current, the rod was magnetized.) Today, this effect is called the *Einstein–de* Ref. 469

* In fact, if one imagines tiny currents moving in circles inside magnets, one gets a unique description for all magnetic fields observed in nature.

* André-Marie Ampère (b. 1775 Lyon, d. 1836 Marseille), French physicist and mathematician. Autodidact, he read the famous *Encyclopédie* as a child; in a life full of personal tragedies, he wandered from maths to chemistry and physics, worked as a school teacher, and published nothing of importance until 1820. Then the discovery of Oersted reached all over Europe: electrical current can deviate magnetic needles. Ampère worked for years on the problem, and in 1826 published the summary of his findings, which lead Maxwell to call him the 'Newton of electricity'. Ampère named and developed many areas of electrodynamics. In 1832, he and his technician also built the first dynamo, or rotative current generator. Of course, the unit of electrical current is named after him.

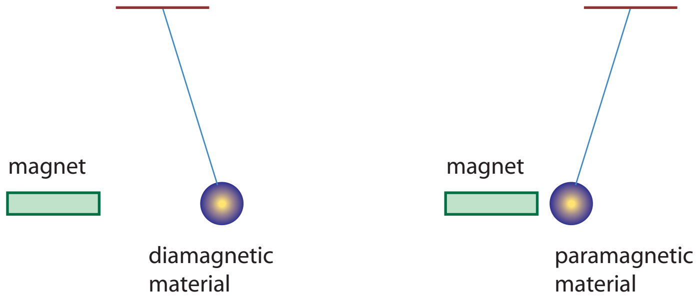


Figure 224 The two basic types of magnetic material behaviour (tested in an inhomogeneous field): diamagnetism and paramagnetism

Haas effect after the two physicists who imagined, measured and explained it.* The effect thus shows that even in the case of a permanent magnet, the magnetic field is due to the internal motion of charges. The size of the effect also shows that the moving particles are electrons. (Twelve years later it became clear that the angular momentum of the electrons responsible for the effect is a mixture of orbital and spin angular momentum; in fact, the electron spin plays a central role in the effect.)

Since magnetism is due to the alignment of microscopic rotational motions, an even more surprising effect can be predicted. Simply rotating a ferromagnetic material* should magnetize it, because the tiny rotating currents would then be aligned along the axis of rotation. This effect has indeed been observed; it is called the *Barnett effect* after its discoverer. Like the Einstein–de Haas effect, the Barnett effect can also be used to determine the gyromagnetic ratio of the electron; thus it also proves that the spins of electrons (usually) play a larger role in magnetism than their orbital angular momentum.

Ref. 470

Page 710

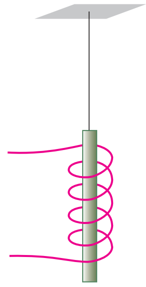


Figure 223 Current makes a metal rods rotate

Magnetic fields

Experiments show that the magnetic field always has a given direction in space, and a magnitude common to all (resting) observers, whatever their orientation. We are tempted to describe the magnetic field by a vector. However, this would be wrong, since a magnetic field does not behave like an arrow when placed before a mirror. Imagine that a system produces a magnetic field directed to the right. You can take any system, a coil, a machine, etc. Now build or imagine a second system that is the exact mirror version of the first: a mirror coil, a mirror machine, etc. The magnetic system produced by the mirror system does not point to the left, as maybe you expected: it still points to the right. (Check by yourself.) In simple

Challenge 901 e

* Wander Johannes de Haas (1878–1960), Dutch physicist. De Haas is best known for two additional magneto-electric effects named after him, the Shubnikov–de Haas effect (the strong increase of the magnetic resistance of bismuth at low temperatures and high magnetic fields) and the de Haas–Van Alphen effect (the diamagnetic susceptibility of bismuth at low temperatures is a periodic function of the magnetic field).

* A ferromagnetic material is a special kind of paramagnetic material that has a permanent magnetization.

words, magnetic fields do *not* behave like arrows.

In other words, it is not completely correct to describe a magnetic field by a vector $\mathbf{B} = (B_x, B_y, B_z)$, as vectors behave like arrows. One also speaks of a *pseudovector*; angular momentum and torque are also examples of such quantities. The precise way is to describe the magnetic field by the quantity*

$$\mathrm{B} = \begin{pmatrix} 0 & -B_z & B_y \\ B_z & 0 & -B_x \\ -B_y & B_x & 0 \end{pmatrix} , \tag{394}$$

called an *antisymmetric tensor*. In summary, *magnetic fields* are defined by the acceleration they impart on moving charges. This acceleration turns out to follow

$$\mathbf{a} = \frac{e}{m}\mathbf{v}\mathrm{B} = \frac{e}{m}\mathbf{v} \times \mathbf{B} \tag{395}$$

a relation which is often called *Lorentz acceleration*, after the important Dutch physicist Hendrik A. Lorentz (b. 1853 Arnhem, d. 1928 Haarlem) who first stated it clearly.* The unit of the magnetic field is called tesla and is abbreviated T. One has $1\,\mathrm{T} = 1\,\mathrm{N\,s/C\,m} = 1\,\mathrm{V\,s/m^2} = 1\,\mathrm{V\,s^2/A\,m}$.

The Lorentz acceleration is the effect at the root of any electric motor. An electric motor is a device that uses a magnetic field as efficiently as possible to accelerate charges flowing in a wire. Through the motion of the charges, the wire is then also moved. Electricity is thus transformed into magnetism and then into motion. The first efficient motor was built back in 1834.

As in the electric case, we need to know how the *strength* of a magnetic field is determined. Experiments such as Oersted's show that the magnetic field is due to moving charges, and that a charge moving with velocity $\mathbf{v}$ produces a field B given by

$$\mathrm{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} q \frac{\mathbf{v} \times \mathbf{r}}{r^3} \quad \text{where} \quad \frac{\mu_0}{4\pi} = 10^{-7}\,\mathrm{N/A^2} . \tag{396}$$

Again, the strange factor $\mu_0/4\pi$ is due to the historical way in which the electrical units were defined. The constant μ_0 is called the *permeability of the vacuum* and is defined by the fraction of newtons per ampere squared given in the formula. It is easy to see that the magnetic field has an intensity given by $\mathbf{vE}/c^2$, where $\mathbf{E}$ is the electric field measured by an observer moving *with* the charge. This is the first hint that magnetism is a relativistic effect.

Challenge 903 e

We note that equation (396) is valid only for small velocities and accelerations. Can you find the general one?

Challenge 904 ny

In 1831, Michael Faraday discovered an additional piece of the puzzle, one that even the great Ampère had overlooked. He found that a *moving* magnet could cause a current

* The quantity B was not called the 'magnetic field' until recently. We follow here the modern, logical definition, which supersedes the traditional one, where B was called the 'magnetic flux density' or 'magnetic induction' and another quantity, **H**, was called – incorrectly, but for over a century – the magnetic field. This quantity **H** will not appear in this walk, but it is important for the description of magnetism in materials.

* Does the definition of magnetic field given here assume a charge speed much lower than that of light?

Challenge 902 ny

flow in an electrical circuit. Magnetism can thus be turned into electricity. This important discovery allowed the production of electrical current flow by generators, so-called *dynamos*, using water power, wind power or steam power. In fact, the first dynamo was built in 1832 by Ampère and his technician. Dynamos started the use of electricity throughout the world. Behind every electrical plug there is a dynamo somewhere.

Additional experiments show that magnetic fields also lead to electric fields when one changes to a moving viewpoint. You might check this on any of the examples of Figures 221 to 232. *Magnetism indeed is relativistic electricity*. Electric and magnetic fields are partly transformed into each other when switching from one inertial reference frame to the other. Magnetic and electrical fields thus behave like space and time, which are also mixed up when changing from one inertial frame to the other. The theory of special relativity thus tells us that there must be a single concept, an *electromagnetic field*, describing them both. Investigating the details, one finds that the electromagnetic field F surrounding charged bodies has to be described by an antisymmetric 4-*tensor*

$$\mathrm{F}^{\mu\nu} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix} \text{ or } \mathrm{F}_{\mu\nu} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix} . \tag{397}$$

Obviously, the electromagnetic field F, and thus every component of these matrices, depends on space and time. The matrices show that electricity and magnetism are two faces of the same effect.* In addition, since electric fields appear only in the topmost row and leftmost column, the expressions show that in everyday life, for small speeds, electricity and magnetism *can* be separated.

The total influence of electric and magnetic fields on fixed or moving charges is then given by the following expression for the relativistic force-acceleration relation $\mathbf{K} = m\mathbf{b}$:

$$\begin{aligned} m\mathbf{b} &= q\mathrm{F}u \quad \text{or} \\ m\,\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} &= q\mathrm{F}^\mu{}_\nu u^\nu \quad \text{or} \\ m\frac{\mathrm{d}}{\mathrm{d}\tau}\begin{pmatrix} \gamma c \\ \gamma v_x \\ \gamma v_y \\ \gamma v_z \end{pmatrix} &= q\begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ E_x/c & 0 & B_z & -B_y \\ E_y/c & -B_z & 0 & B_x \\ E_z/c & B_y & -B_x & 0 \end{pmatrix}\begin{pmatrix} \gamma c \\ \gamma v_x \\ \gamma v_y \\ \gamma v_z \end{pmatrix} \quad \text{or} \\ W = q\mathbf{E}\mathbf{v} \quad &\text{and} \quad d\mathbf{p}/\mathrm{d}t = q(\mathbf{E} + \mathbf{v}\times\mathbf{B}) \;, \end{aligned} \tag{398}$$

which show how the work W and the three-force $dp/\mathrm{d}t$ depend on the electric and magnetic fields. All four expressions describe the same content; the simplicity of the first one is the reason for the involved matrices (397) of the electromagnetic field. In fact, the extended *Lorentz relation* (398) is the *definition* of the electromagnetic field, since the field is defined as that 'stuff' which accelerates charges. In particular, all devices that put charges

* Actually, the expression for the field contains everywhere the expression $1/\sqrt{\mu_0\varepsilon_0}$ instead of the speed of light c. We will explain the reason for this substitution shortly.

into motion, such as batteries and dynamos, as well as all devices that are put into motion by flowing charges, such as electric motors and muscles, are described by this relation. That is why this relation is usually studied, in simple form, already in school. The Lorentz relation describes all cases in which the motion of objects can be seen by the naked eye or felt by our senses, such as the movement of an electrical motor in a high speed train, in an elevator and in a dental drills, the motion of the picture generating electron beam in a television tube, or the travelling of an electrical signal in a cable and in the nerves of the body. Ref. 471, Ref. 472

In equation (398) it is understood that one sums over indices that appear twice. The electromagnetic field tensor F is an *antisymmetric* 4-tensor. (Can you write down the relation between $F^{\mu\nu}$, $F_{\mu\nu}$ and $F^{\mu}{}_{\nu}$?) Challenge 905 ny Like any such tensor, it has two invariants, i.e., two deduced properties that are the same for every observer: the expression $B^2 - E^2/c^2 = \frac{1}{2}\mathrm{tr}\,F^2$ and the product $4\mathbf{EB} = -c\ \mathrm{tr}\,F^*F$. (Can you confirm this, using the definition of trace as the sum of the diagonal elements?) Challenge 906 n

The first invariant expression turns out to be the Lagrangian of the electromagnetic field. It is a scalar and implies that if E is larger, smaller, or equal to cB for one observer, it also is for all other observers. The second invariant, a pseudoscalar, describes whether the angle between the electric and the magnetic field is acute or obtuse for all observers.*

The application of electromagnetic effects to daily life has opened up a whole new world that did not exist before. Electrical light, electric motors, radio, telephone, X-rays, television and computers have changed human life completely in less than one century. For example, the installation of electric lighting in city streets has almost eliminated the previously so common night assaults. These and all other electrical devices exploit the fact that charges can flow in metals and, in particular, that electromagnetic energy can be transformed

- into mechanical energy – as used in loudspeakers, motors, piezo crystals;
- into light – as in lamps and lasers;
- into heat – as in ovens and tea pots;
- into chemical effects – as in hydrolysis, battery charging and electroplating;
- into coldness – as in refrigerators and Peltier elements;
- into radiation signals – as in radio and television;
- into stored information – as in magnetic records and in computers.

* There is in fact a third Lorentz invariant, much less known. It is specific to the electromagnetic field and is a combination of the field and its vector potential:

$$\begin{aligned}\kappa_3 &= \frac{1}{2}A_\mu A^\mu F_{\rho\nu}F^{\nu\rho} - 2A_\rho F^{\rho\nu}F_{\nu\mu}A^\mu \\ &= (\mathbf{AE})^2 + (\mathbf{AB})^2 - |\mathbf{A}\times\mathbf{E}|^2 - |\mathbf{A}\times\mathbf{B}|^2 + 4\frac{\varphi}{c}(\mathbf{AE}\times\mathbf{B}) - (\frac{\varphi}{c})^2(E^2+B^2)\ . \end{aligned} \quad (399)$$

Ref. 475 This expression is Lorentz (but not gauge) invariant; knowing it can help clarify unclear issues, such as the lack of existence of waves in which the electric and magnetic fields are parallel. Indeed, for plane monochromatic waves all three invariants *vanish* in the Lorentz gauge. Also the quantities $\partial_\mu J^\mu$, $J_\mu A^\mu$ and $\partial_\mu A^\mu$ are Lorentz invariants. (Why?) Challenge 907 n The latter, the frame independence of the divergence of the four-potential, reflects the invariance of gauge choice. The gauge in which the expression is set to zero is called the *Lorentz gauge*.

How motors prove relativity to be right

> The only mathematical operation I performed in my life was to turn the handle of a calculator.
>
> Michael Faraday

All electric motors are based on the result that electric currents interact with magnetic fields. The simplest example is the attraction of two wires carrying parallel currents. This observation alone, made in 1820 by Ampère, is sufficient to make motion larger than a certain maximal speed impossible. Ref. 477

The argument is beautifully simple. We change the original experiment and imagine two long, electrically charged rods of mass m, moving in the same direction with velocity v and separation d. An observer moving with the rods would see an electrostatic repulsion between the rods given by Challenge 908 e

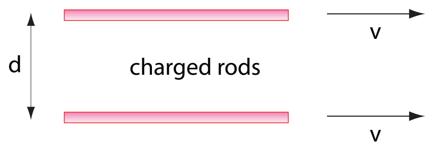


Figure 225 The relativistic aspect of magnetism

$$ma_e = -\frac{1}{4\pi\varepsilon_0}\frac{2\lambda^2}{d} \qquad (400)$$

where λ is the charge per length of the rods. A second, *resting* observer sees two effects: the electrostatic repulsion and the attraction discovered by Ampère. The second observer therefore observes Challenge 909 e

$$ma_{em} = -\frac{1}{4\pi\varepsilon_0}\frac{2\lambda^2}{d} + \frac{\mu_0}{2\pi}\frac{\lambda^2 v^2}{d} \ . \qquad (401)$$

It is easy to check that the second observer sees a repulsion, as does the first one, only if

$$v^2 < \frac{1}{\varepsilon_0\mu_0} \ . \qquad (402)$$

This maximum speed, with a value of 0.3 GM/s, is thus valid for any object carrying charges. But *all* everyday objects contain charges: there is thus a maximum speed for matter.

Challenge 910 ny Are you able to extend the argument for a maximum speed to neutral particles as well? We will find out more on this limit velocity, which we know already, in a minute.

Another argument for magnetism as a relativistic effect is the following. In a wire with electrical current, the charge is zero for an observer at rest with respect to the wire. The reason is that the charges enter and exit the wire at the same time for that observer. Now imagine an observer who flies along the wire. The entrance and exit events do not occur simultaneously any more; the wire is charged for a moving observer. (The charge depends on the direction of the observer's motion.) In other words, if the observer himself were charged, he would experience a force. Moving charges experience forces from current-carrying wires. This is exactly why magnetic fields were introduced: they only produce forces on *moving* charges. In short, current carrying wires are surrounded by magnetic fields.

In summary, electric effects are due to flow of electric charges and to electric fields;

Table 45 Voltage values observed in nature

Observation	Voltage
Smallest measured voltage	0.1 pV
Human nerves	70 mV
Voltaic cell ('battery')	1.5 V
Mains in households	230(15) V
Electric eel	100 to 600 V
Sparks when rubbing a polymer pullover	1 kV
Electric fence	0.7 to 10 kV
Colour television tube	30 kV
X-ray tube	10 to 200 kV
Electron microscopes	0.5 kV to 3 MV
Stun gun	65 to 600 kV
Lightning stroke	10 to 100 MV
Record accelerator voltage	1 TV
Highest voltage possible in nature	$1.5 \cdot 10^{27}$ V

magnetism is due to *moving* electric charges. It is *not* due to magnetic charges.* The strength of magnetism, used in any running electric motor, proves relativity right: there is a maximum speed in nature. Both electric and magnetic fields carry energy and momentum. They are two faces of the same coin. However, our description of electromagnetism is not complete yet: we need the final description of the way charges *produce* an electromagnetic field.

Curiosities and fun challenges about things electric and magnetic

Et facta mirari et intellectua assequi.

Augustine

Before we study the motion of an electromagnetic field in detail, let's have some fun with electricity.

▪ If even knocking on a wooden door is an electric effect, we should be able to detect Challenge 912 ny fields when doing so. Can you devise an experiment to check this?

▪ Birds come to no harm when they sit on unprotected electricity lines. Nevertheless, one almost never observes any birds on tall, high voltage lines of 100 kV or more, which Challenge 913 n transport power across longer distances. Why?

▪ How can you distinguish a magnet from an unmagnetized metal bar of the same size Challenge 914 n and material, using no external means?

▪ How do you wire up a light bulb to the mains and three switches so that the light can Challenge 915 n be switched on at any of the switches and off at any other switch? And for four switches?

Page 545 Challenge 911 ny * 'Electrons move in metal with a speed of about 1 μm/s; thus if I walk with the same speed along a cable carrying a constant current, I should not be able to sense any magnetic field.' What is wrong with this argument?

Nobody will take a physicist seriously who is able to write Maxwell's equations but cannot solve this little problem.

▪ The first appliances built to generate electric currents were large rubbing machines. Then, in 1799 the Italian scientist Alessandro Volta (1745–1827) invented a new device to generate electricity and called it a *pile*; today it is called a (voltaic) cell or, less correctly, a *battery*. Voltaic cells are based on chemical processes, provide much more current, work in all weathers, and are smaller and easier to handle than electrostatic machines. The invention of the battery changed the investigation of electricity so profoundly that Volta became world famous. At last, a simple source of electricity was available for use in experiments; unlike rubbing machines, piles are compact, work in all weather conditions and make no noise.

An apple or a potato with a piece of copper and one of zinc inserted is one of the simplest possible voltaic cells. It provides about 1 V of electrical tension and can be used to run digital clocks or to produce clicks in headphones. Volta was also the discoverer of the charge law $q = CU$ of capacitors (C being the capacity, and U the voltage) and the inventor of the high sensitivity capacitor electroscope. A modest man, nevertheless, the unit of electrical potential, or 'tension', as Volta used to call it, was deduced from his name. A 'battery' is a large number of voltaic cells; the term was taken from an earlier, almost purely military use.* A batteries in a mobile phone is just an elaborated replacement for a number of apples or potatoes.

▪ A PC or a telephone can communicate without wires, by using radio waves. Why are these and other electrical appliances not able to obtain their *power* via radio waves, thus
Challenge 917 n eliminating power cables?

▪ Objects that are not right–left symmetric are called *chiral*, from the Greek word for
Challenge 918 n 'hand'. Can you make a mirror that does not exchange left and right? In two different ways?

▪ A Scotch tape roll is a dangerous device. Pulling the roll quickly leads to light emission (through triboluminescence) and to small sparks. It is suspected that several explosions in mines were triggered when such a spark ignited a combustible gas mixture.

▪ Take an envelope, wet it and seal it. After letting it dry for a day or more, open it in the dark. At the place where the two sides of paper are being separated from each other,
Challenge 919 n the envelope glows with a blue colour. Why? Is it possible to speed up the test using a hair dryer?

▪ Electromagnetism is full of surprises and offers many effects that can be reproduced at home. The internet is full of descriptions of how to construct Tesla coils to produce sparks, coil guns or rail guns to shoot objects, electrostatic machines to make your hair stand on end, glass spheres with touch-sensitive discharges and much more. If you like experiments, just search for these terms.

▪ A high voltage can lead to current flow through air, because air becomes conductive in high electric fields. In such discharges, air molecules are put in motion. As a result, one can make objects that are attached to a pulsed high tension source lift up in the air, if one optimizes this air motion so that it points downwards everywhere. The high tension is thus effectively used to accelerate ionized air in one direction and, as a result, an object

* A pile made of sets of a zinc plate, a sheet of blotting paper soaked with salt water and a copper coin is
Challenge 916 ny easily constructed at home.

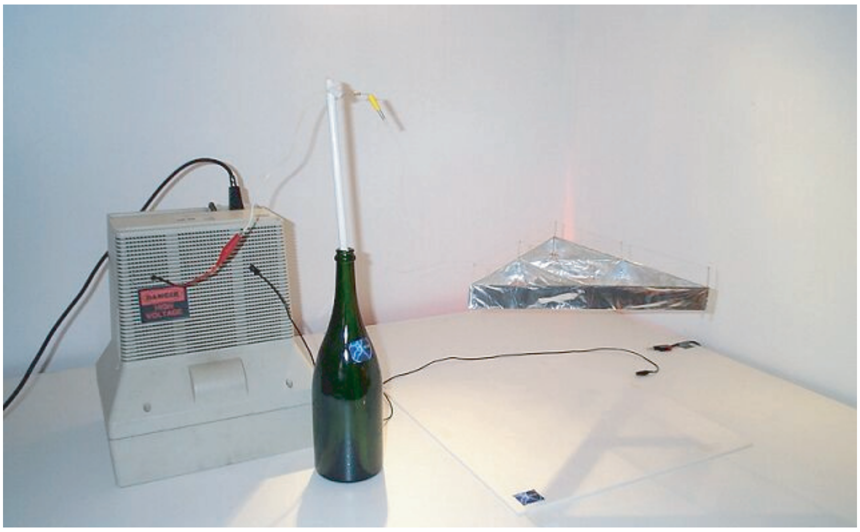


Figure 226 Lifting a light object – covered with aluminium foil – using high a tension discharge (© Jean-Louis Naudin at http://www.jlnlabs.org)

will move in the opposite direction, using the same principle as a rocket. An example is shown in Figure 226, using the power supply of a PC monitor. (Watch out: danger!) Numerous websites explain how to build these so-called lifters at home; in Figure 226, the bottle and the candle are used as high voltage insulator to keep one of the two thin high voltage wires (not visible in the photograph) high enough in the air, in order to avoid discharges to the environment or to interfere with the lifter's motion. Unfortunately, the majority of websites – not all – give incorrect or confused explanations of the phenomenon. These websites thus provide a good challenge for one to learn to distinguish fact from speculation.

Challenge 920 e

▪ The electric effects produced by friction and by liquid flow are usually small. However, in the 1990s, a number oil tankers disappeared suddenly. The sailors had washed out the oil tanks by hosing sea water onto the tank walls. The spraying led to charging of the tank; a discharge then led to the oil fumes in the tank igniting. This led to an explosion and subsequently the tankers sank. Similar accidents also happen regularly when chemicals are moved from one tank to another.

▪ Rubbing a plastic spoon with a piece of wool charges it. Such a charged spoon can be used to extract pepper from a salt–pepper mixture by holding the spoon over the mixture. Why?

Challenge 921 n

▪ When charges move, they produce a magnetic field. In particular, when ions inside the Earth move due to heat convection, they produce the Earth's magnetic field. When the ions high up in the stratosphere are moved by solar wind, a geomagnetic storm appears; its field strength can be as high as that of the Earth itself. In 2003, an additional mechanism was discovered. When the tides move the water of the oceans, the ions in the salt water produce a tiny magnetic field; it can be measured by highly sensitive magnetometers in satellites orbiting the Earth. After two years of measurements from a small satellite it was possible to make a beautiful film of the oceanic flows. Figure 227 gives an

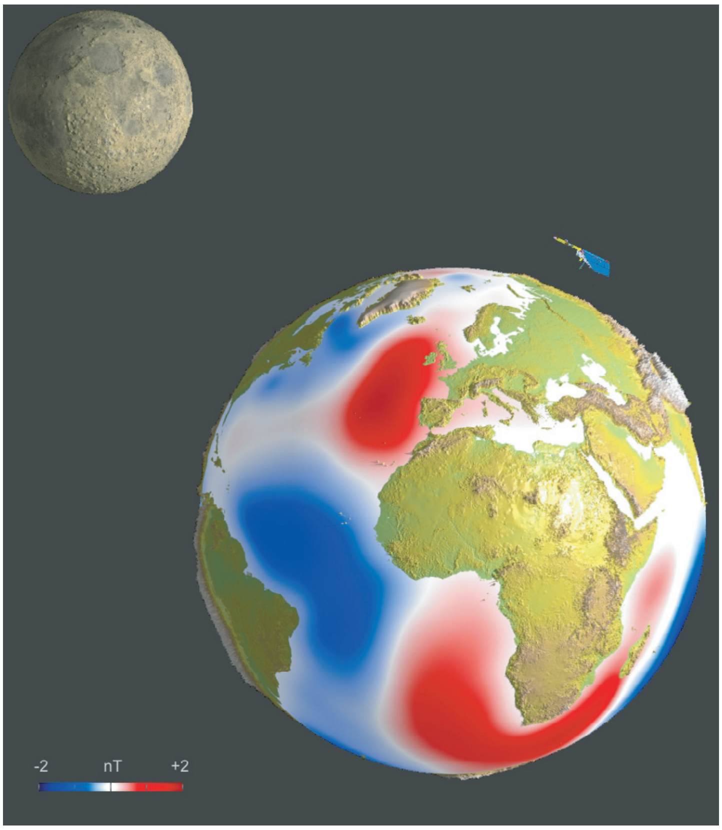


Figure 227 The magnetic field due to the tides

Ref. 473 impression.

▪ The names electrode, electrolyte, ion, anode and cathode were suggested by William Whewell (1794–1866) on demand of Michael Faraday; Faraday had no formal education and asked his friend Whewell to form two Greek words for him. For anode and cathode, Whewell took words that literally mean 'upward street' and 'downward street'. Faraday then popularized these terms, like the other words mentioned above.

▪ The shortest light pulse produced so far had a length of 100 as. To how many wavelengths of green light would that correspond? Challenge 922 n

▪ Why do we often see shadows of houses and shadows of trees, but never shadows of the electrical cables hanging over streets? Challenge 923 n

▪ How would you measure the speed of the tip of a lightning bolt? What range of values do you expect? Challenge 924 n

Ref. 527 ▪ One of the simplest possible electric motors was discovered by Faraday in 1831. A magnet suspended in mercury will start to turn around its axis if a current flows through it. (See Figure 228.) In addition, when the magnet is forced to turn, the device (often also called Barlow's wheel) also works as a current generator; people have even tried to generate domestic current with such a system! Can you explain how it works? Challenge 925 n

The modern version of this motor makes use of a battery, a wire, a conductive

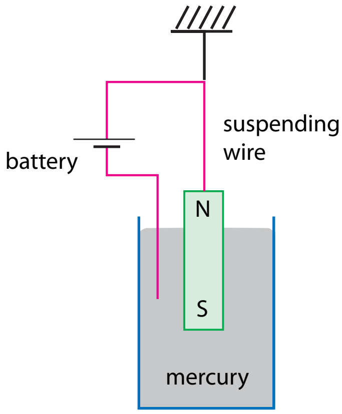


Figure 228 A unipolar motor

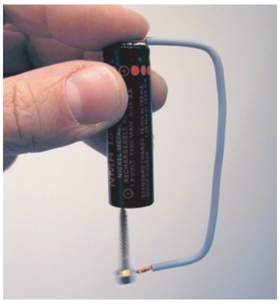


Figure 229 The simplest motor (© Stefan Kluge)

samarium–cobalt magnet and a screw. The result is shown in Figure 229.

▪ The magnetic field of the Earth is much higher than that of other planets because of the Earth's Moon. The field has a dipole strength of $7.8 \cdot 10^{22}$ A m^2. It shields us from lethal solar winds and cosmic radiation particles. Today, a lack of magnetic field would lead to high radiation on sunny days; but in the past, its lack would have prevented the evolution of the human species. We owe our existence to the magnetic field. Ref. 544

▪ The ionosphere around the Earth has a resonant frequency of 7 Hz; for this reason any apparatus measuring low frequencies always gets a strong signal at this value. Can you give an explanation of the frequency? Challenge 926 n

▪ The Sun is visible to the naked eye only up to a distance of 50 light years. Is this true? Challenge 927 ny

▪ At home, electricity is mostly used as alternating current. In other words, no electrons actually flow through cables; as the drift speed of electrons in copper wires is of the order of 1 µm/s, electrons just move back and forward by 20 nm. Nothing flows in or out of the cables! Why do the electricity companies require a real flow of money in return, instead of being satisfied with a back and forth motion of money? Page 545 Challenge 928 e

▪ Comparing electricity with water is a good way of understanding electronics. Figure 230 shows a few examples that even a teenager can use. Can you fill in the correspondence for the coil, and thus for a transformer? Challenge 929 n

▪ Do electrons and protons have the same charge? Experiments show that the values are equal to within at least twenty digits. How would you check this? Challenge 930 ny

▪ Charge is also velocity-independent. How would you check this? Challenge 931 ny

▪ Magnets can be used, even by school children, to climb steel walls. Have a look at the http://www.physicslessons.com/TPNN.htm website.

▪ Extremely high magnetic fields have strange effects. At fields of 10^{10} T, vacuum becomes effectively birefringent, photons can split and coalesce, and atoms get squeezed. Hydrogen atoms, for example, are estimated to get two hundred times narrower in one direction. Fortunately, these conditions exist only in specific neutron stars, called magnetars.

▪ A good way to make money is to produce electricity and sell it. In 1964, a completely

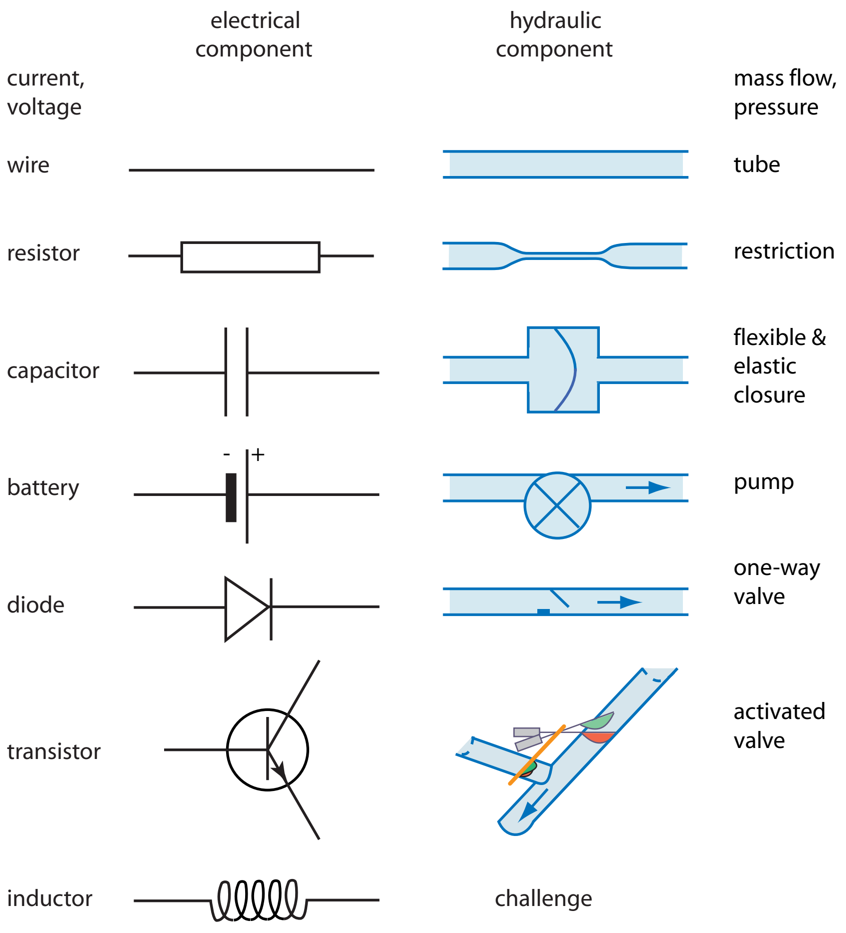


Figure 230 The correspondence of electronics and water flow

new method was invented by Fletcher Osterle. The method was presented to a larger public in a beautiful experiment in 2003. One can take a plate of glass, add a conducting layers on each side, and then etch a few hundred thousand tiny channels through the plate, each around 15 μm in diameter. When water is made to flow through the channels, a current is generated. The contacts at the two conducting plates can be used like battery contacts. Ref. 474

This simple device uses the effect that glass, like most insulators, is covered with a charged layer when it is immersed in a liquid. Can you imagine why a current is generated? Unfortunately, the efficiency of electricity generation is only about 1%, making the method much less interesting than a simple blade wheel powering a dynamo. Challenge 932 n

The description of electromagnetic field evolution

In the years between 1861 and 1865, taking in the details of all the experiments known to him, James Clerk Maxwell produced a description of electromagnetism that forms one

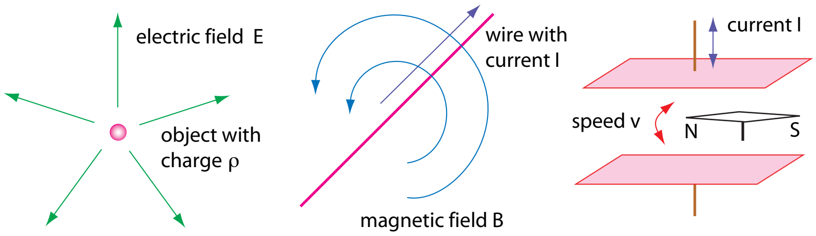


Figure 231 The first of Maxwell's equations

of the pillars of physics.* Maxwell took all the experimental results and extracted their common basic principles, as shown in Figures 231 and 232. Twenty years later, Heaviside and Hertz extracted the main points of Maxwell ideas, calling their summary *Maxwell's theory of the electromagnetic field*. It consists of two equations (four in the non-relativistic case).

Page 1104

The first equation is the precise statement that electromagnetic fields *originate at charges*, and nowhere else. The corresponding equation is variously written*

$$d\mathrm{F} = j\sqrt{\frac{\mu_0}{\varepsilon_0}} \quad \text{or}$$

$$d^\nu \mathrm{F}_{\mu\nu} = j^\mu \sqrt{\frac{\mu_0}{\varepsilon_0}} \quad \text{or} \tag{403}$$

$$(\partial_t/c, -\partial_x, -\partial_y, -\partial_z)\begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix} = \sqrt{\frac{\mu_0}{\varepsilon_0}}(\rho, j_x/c, j_y/c, j_z/c) \text{ or}$$

$$\nabla \mathbf{E} = \frac{\rho}{\varepsilon_0} \quad \text{and} \quad \nabla \times \mathbf{B} - \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} = \mu_0 \mathbf{j}\ .$$

Each of these four equivalent ways to write the equation makes a simple statement: *electrical charge carries the electromagnetic field*. This statement, including its equations, are equivalent to the three basic observations of Figure 231. It describes Coulomb's relation, Ampère's relation, and the way changing currents induce magnetic effects, as you may want to check for yourself.

Challenge 933 ny

The second result by Maxwell is the precise description of how changing electric fields create magnetic fields, and vice versa. In particular, an electric field can have vortices only when there is a changing magnetic field. In addition it expresses the observation that in

* James Clerk Maxwell (b. 1831 Edinburgh, d. 1879 Cambridge), Scottish physicist; founded electromagnetism by unifying electricity and magnetism theoretically, as described in this chapter. His work on thermodynamics forms the second pillar of his activity. In addition, he studied the theory of colours and developed the now standard horseshoe colour diagram; he was one of the first people to make a colour photograph. He is regarded by many as the greatest physicist ever. Both 'Clerk' and 'Maxwell' both were his family names.

* Maxwell generalized this equation to cases where the charges are not surrounded by vacuum, but located inside matter. We will not explore these situations in our walk because, as we will see during our mountain ascent, the apparently special case of vacuum in fact describes all of nature.

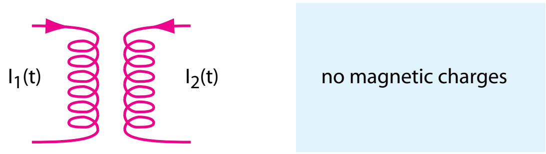


Figure 232 The second of Maxwell's equations

nature there are no magnetic charges, i.e. that magnetic fields have no sources. All these results are described by the relation variously written

$$
\begin{gathered}
d\,{}^*\mathrm{F} = 0 \quad \text{with} \quad {}^*\mathrm{F}^{\rho\sigma} = \frac{1}{2}\varepsilon^{\rho\sigma\mu\nu}\mathrm{F}_{\mu\nu} \quad \text{or} \\
\varepsilon_{\mu\nu\rho}\partial_\mu \mathrm{F}_{\nu\rho} = \partial_\mu \mathrm{F}_{\nu\rho} + \partial_\nu \mathrm{F}_{\rho\mu} + \partial_\rho \mathrm{F}_{\mu\nu} = 0 \quad \text{or} \\
\begin{pmatrix} \gamma\frac{1}{c}\partial_t \\ \gamma\partial_x \\ \gamma\partial_y \\ \gamma\partial_z \end{pmatrix}
\begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z/c & E_y/c \\ -B_y & E_z/c & 0 & -E_x/c \\ -B_z & -E_y/c & E_x/c & 0 \end{pmatrix}
= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{or} \\
\nabla \mathbf{B} = 0 \quad \text{and} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \,.
\end{gathered} \tag{404}
$$

The relation expresses the *lack of sources for the dual field tensor*, usually written *F: there are no magnetic charges, i.e. no magnetic monopoles in nature. In practice, this equation is always needed together with the previous one. Can you see why? Challenge 934 ny

We now have a system as organized as the expression $a = GM/r$ or as Einstein's field equations for gravitation. Together with Lorentz' evolution equation (398), which describes how charges move given the motion of the fields, Maxwell's evolution equations (403) and (404) describe *all* electromagnetic phenomena occurring on everyday scales, from mobile phones, car batteries, to personal computers, lasers, lightning, holograms and rainbows.

We will not study many applications of the equations but will continue directly towards our aim to understand the connection to everyday motion and to the motion of light. In fact, the electromagnetic field has an important property that we mentioned right at the start: the field itself itself can move.

Colliding charged particles

A simple experiment clarifies the properties of electromagnetic fields defined above. When two charged particles collide, their total momentum is *not* conserved.

Imagine two particles of identical mass and identical charge just after a collision, when they are moving away from one another. Imagine also that the two masses are large, so that the acceleration due to their electrical repulsion is small. For an observer at the centre of gravity of the two, each particle feels an acceleration from the electric field of the other. The electric field E is given by the so-called *Heaviside formula* Challenge 935 ny

$$E = \frac{q(1-v^2/c^2)}{4\pi e_0 r^2} \; . \tag{405}$$

In other words, the total system has a vanishing total momentum.

Take a second observer, moving with respect to the first with velocity v, so that the first charge will be at rest. Expression (405) leads to two *different* values for the electric fields, one at the position of each particle. In other words, the system of the two particles is not in inertial motion, as we would expect; the total momentum is not conserved. Where did it go?

Ref. 481

Challenge 936 n

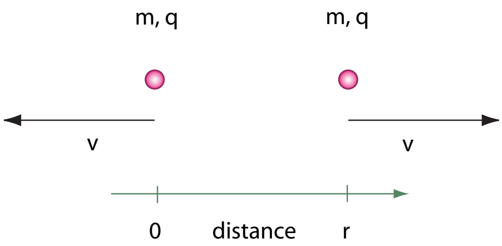


Figure 233 Charged particles after a collision

This at first surprising effect has been put in the form of a theorem by Van Dam and Wigner. They showed that for a system of particles interacting at a distance the total particle energy–momentum cannot remain constant in all inertial frames.

Ref. 482

The total momentum of the system is conserved only because the electromagnetic field itself also carries momentum. If electromagnetic fields have momentum, they are able to *strike* objects and to be struck by them. As we will show below, light is also an electromagnetic field. Thus we should be able to move objects by shining light on to them. We should even be able to suspend particles in mid air by shining light on to them from below. Both predictions are correct, and some experiments will be presented shortly.

We conclude that any sort of field leading to particle interactions must carry energy and momentum, as the argument applies to all such cases. In particular, it applies to nuclear interactions. Indeed, in the second part of our mountain ascent we will even find an additional result: all fields are themselves composed of particles. The energy and momentum of fields then become an obvious state of affairs.

The gauge field: the electromagnetic vector potential

The study of moving fields is called *field theory* and electrodynamics is the prime example. (The other classical example is fluid dynamics; moving electromagnetic fields and moving fluids are very similar mathematically.) Field theory is a beautiful topic; field lines, equipotential lines and vortex lines are some of the concepts introduced in this domain. They fascinate many.* However, in this mountain ascent we keep the discussion focused on motion.

We have seen that fields force us to extend our concept of motion. Motion is not only the change in state of objects and of space-time, but also the *change in state of fields*. We therefore need, also for fields, a complete and precise description of their state. The observations with amber and magnets have shown us that *fields possess energy and momentum*. They can impart it to particles. The experiments with motors have shown that objects can

Challenge 937 n

* What is the relation, for static fields, between field lines and (equi-) potential surfaces? Can a field line cross a potential surface twice? For more details on topics such as these, see the *free* textbook by Bo Thidé, *Electromagnetic Field Theory*, on his http://www.plasma.uu.se/CED/Book website. And of course, in English, have a look at the texts by Schwinger and by Jackson.

Ref. 439

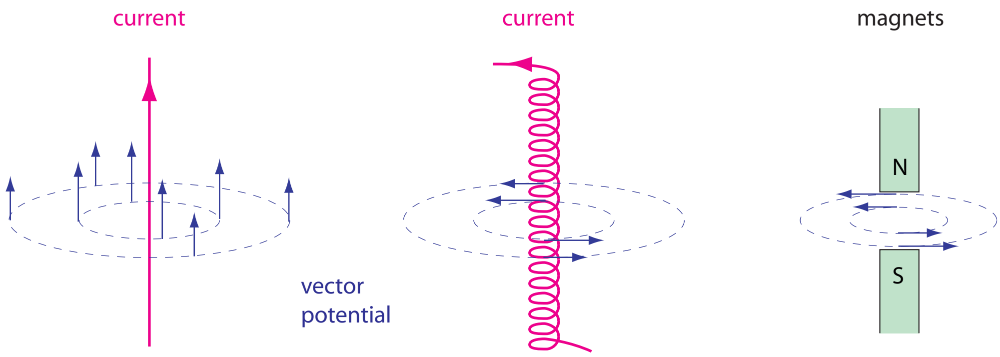


Figure 234 Vector potentials for selected situations

add energy and momentum to fields. We therefore have to define a *state function* which allows us to define energy and momentum for electric and magnetic fields. Since electric and magnetic fields transport energy, their motion follows the speed limit in nature.

Maxwell defined the state function in two standard steps. The first step is the definition of the *(magnetic) vector potential*, which describes the momentum per charge that the

Ref. 478 field provides:

$$\mathbf{A} = \frac{\mathbf{p}}{q} . \tag{406}$$

When a charged particle moves through a magnetic potential $\mathbf{A}(\mathbf{x})$, its momentum changes by $q\Delta\mathbf{A}$; it changes by the difference between the potential values at the start and end points, multiplied by its charge. Owing to this definition, the vector potential has the property that

$$\mathrm{B} = \nabla \times \mathbf{A} = \text{curl } \mathbf{A} \tag{407}$$

i.e. that the magnetic field is the curl of the magnetic potential. The curl is called the *rotation*, abbreviated rot in most languages. The curl (or rotation) of a field describes, for each point of space, the direction of the local, imagined axis of rotation, as well as (twice) the rotation speed around that axis. For example, the curl for the velocities of a rotating

Challenge 938 ny solid body is everywhere 2ω, or twice the angular velocity.

Ref. 479 The vector potential for a long straight current-carrying wire is parallel to the wire; it

Challenge 939 d has the magnitude

$$A(r) = -\frac{\mu_0 I}{4\pi} \ln \frac{r}{r_0} , \tag{408}$$

which depends on the radial distance r from the wire and an integration constant r_0. This expression for the vector potential, pictured in Figure 234, shows how the moving current produces a linear momentum in the (electro-) magnetic field around it. In the case of a solenoid, the vector potential 'circulates' around the solenoid. The magnitude obeys

$$A(\mathrm{r}) = -\frac{\Phi}{4\pi} \frac{1}{r} , \tag{409}$$

where Φ is the magnetic flux inside the solenoid. We see that, in general, the vector potential is *dragged along* by moving charges. The dragging effect decreases for larger distances. This fits well with the image of the vector potential as the momentum of the electromagnetic field.

This behaviour of the vector potential around charges is reminiscent of the way honey is dragged along by a spoon moving in it. In both cases, the dragging effect decreases with distance. However, the vector potential, unlike the honey, does *not* produce any friction that slows down charge motion. The vector potential thus behaves like a frictionless liquid.

Inside the solenoid, the magnetic field is constant and uniform. For such a field B we
Challenge 940 ny
find the vector potential

$$\mathbf{A}(\mathrm{r}) = -\frac{1}{2}\mathbf{B} \times \mathbf{r} \,. \tag{410}$$

In this case, the magnetic potential thus increases with increasing distance from the origin.* In the centre of the solenoid, the potential vanishes. The analogy of the dragged honey gives exactly the same behaviour.

However, there is a catch. The magnetic potential is *not* defined uniquely. If $\mathbf{A}(\mathbf{x})$ is a vector potential, then the different vector potential

$$\mathbf{A}'(\mathbf{x}) = \mathbf{A}(\mathbf{x}) + \operatorname{grad} \Lambda \,, \tag{411}$$

where $\Lambda(t, \mathbf{x})$ is some scalar function, is *also* a vector potential for the same situation. (The magnetic field B stays the same, though.) Worse, can you confirm that the corres-
Challenge 941 ny
ponding (absolute) momentum values also change? This unavoidable ambiguity, called *gauge invariance*, is a central property of the electromagnetic field. We will explore it in more detail below.

Not only the momentum, but also the energy of the electromagnetic field is defined ambiguously. Indeed, the second step in the specification of a state for the electromagnetic
Ref. 478
field is the definition of the *electric potential* as the energy U per charge:

$$\varphi = \frac{U}{q} \tag{412}$$

In other words, the potential $\varphi(\mathbf{x})$ at a point $\mathbf{x}$ is the energy needed to move a unit charge to the point $\mathbf{x}$ starting from a point where the potential vanishes. The potential energy is thus given by $q\varphi$. From this definition, the electric field $\mathbf{E}$ is simply the *change* of the potential with position corrected by the time dependence of momentum, i.e.

$$\mathbf{E} = -\nabla\varphi - \frac{\partial}{\partial t}\mathbf{A} \,, \tag{413}$$

Obviously, there is a freedom in the choice of the definition of the potential. If $\varphi(\mathbf{x})$ is a

* This is only possible as long as the field is constant; since all fields drop again at large distances – because the energy of a field is always finite – also the vector potential drops at large distances.

possible potential, then

$$\varphi'(\mathbf{x}) = \varphi(\mathbf{x}) - \frac{\partial}{\partial t}\mathbf{\Lambda} \qquad (414)$$

is also a potential function for the same situation. This freedom is the generalization of the freedom to define energy up to a constant. Nevertheless, the electric field **E** remains the same for all potentials.

Ref. 478
Challenge 942 ny
To be convinced that the potentials really are the energy and momentum of the electromagnetic field, we note that for a moving charge we have

$$\begin{aligned} \frac{d}{\mathrm{d}t}\left(\frac{1}{2}mv^2 + q\varphi\right) &= \frac{\partial}{\partial t}q(\varphi - \mathbf{v}\mathbf{A}) \\ \frac{d}{\mathrm{d}t}(m\mathbf{v} + q\mathbf{A}) &= -\nabla q(\varphi - \mathbf{v}\mathbf{A}) \;, \end{aligned} \qquad (415)$$

which show that the changes of generalized energy and momentum of a particle (on the left-hand side) are due to the change of the energy and momentum of the electromagnetic field (on the right-hand side).*

In relativistic 4-vector notation, the energy and the momentum of the field appear together in one quantity. The state function of the electromagnetic field becomes

$$A^{\mu} = (\varphi/c, \mathbf{A}) \;. \qquad (416)$$

It is easy to see that the description of the field is complete, since we have

$$\mathrm{F} = d\,A \quad \text{or} \quad \mathrm{F}^{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu} \;, \qquad (417)$$

which means that the electromagnetic field F is completely specified by the 4-potential A. But as just said, the 4-potential itself is *not* uniquely defined. Indeed, any other gauge field A' is related to A by the *gauge transformation*

$$A'^{\,\mu} = A^{\mu} + \partial^{\mu}\Lambda \qquad (418)$$

where $\Lambda = \Lambda(t, x)$ is any arbitrarily chosen scalar field. The new field A' leads to the *same* electromagnetic field, and to the same accelerations and evolutions. The gauge 4-field A is thus an *overdescription* of the physical situation as several *different* gauge fields correspond to the *same* physical situation. Therefore we have to check that all measurement results are independent of gauge transformations, i.e. that all observables are gauge invariant quantities. Such gauge invariant quantities are, as we just saw, the fields F and *F, and in general all classical quantities. We add that many theoretical physicists use the term 'electromagnetic field' loosely for both the quantities $F^{\mu\nu}$ and A_{μ}.

There is a simple image, due to Maxwell, to help overcoming the conceptual difficulties of the vector potential. It turns out that the closed line integral over A_{μ} is gauge invariant,
Challenge 943 e
because

* This connection also shows why the expression $P^{\mu} - qA^{\mu}$ appears so regularly in formulae; indeed, it plays a central role in the quantum theory of a particle in the electromagnetic field.

$$\oint A_\mu \mathrm{d}x^\mu = \oint (A_\mu + \partial_\mu \Lambda)\mathrm{d}x^\mu = \oint A'_\mu \mathrm{d}x^\mu \; . \tag{419}$$

In other words, if we picture the vector potential as a quantity allowing one to associate a number to a tiny ring at each point in space, we get a good, gauge invariant picture of the vector potential.*

Now that we have defined a state function that describes the energy and momentum of the electromagnetic field, let us look at what happens in more detail when electromagnetic fields move.

Energy, linear and angular momentum of the electromagnetic field

The description so far allows us to write the *total* energy E_{nergy} of the electromagnetic field as

$$E_{\text{nergy}} = \frac{1}{8\pi} \int \varepsilon_0 \mathbf{E}^2 + \frac{\mathbf{B}^2}{\mu_0} \, dV \; . \tag{420}$$

Energy is quadratic in the fields.

For the total linear momentum one obtains

$$\mathbf{P} = \frac{\varepsilon_0}{4\pi} \int \mathbf{E} \times \mathbf{B} \, dV \; . \tag{421}$$

For the total angular momentum one has

$$\mathbf{P} = \frac{\varepsilon_0}{4\pi} \int \mathbf{E} \times \mathbf{A} \, dV \; , \tag{422}$$

where **A** is the vector potential.

The Lagrangian of electromagnetism

The motion of charged particles and the motion of the electromagnetic field can also be described using a Lagrangian instead of using the three equations given above. It is not hard to see that the action S_{CED} for a particle in classical electrodynamics can be
Challenge 944 ny symbolically defined by**

$$S_{\text{CED}} = -mc^2 \int \mathrm{d}\tau - \tfrac{1}{4\mu_0} \int \mathrm{F} \wedge * \mathrm{F} - \int j \wedge A \; , \tag{423}$$

Ref. 480 * In the second part of the text, on quantum mechanics, we will see that the exponent of this expression, namely $\exp(iq \oint A_\mu \mathrm{d}x^\mu)/\hbar$, usually called the *phase factor*, can indeed be directly observed in experiments.

** The product described by the symbol $\wedge$, 'wedge' or 'hat', has a precise mathematical meaning, defined for this case in equation (424). Its background, the concept of *(mathematical) form*, carries us too far from
Ref. 483 our walk.

which in index notation becomes

$$S_{\text{CED}} = -mc \int_{-\infty}^{\infty} \sqrt{\eta_{\mu\nu} \frac{\mathrm{d}x_n^\mu(s)}{\mathrm{d}s} \frac{\mathrm{d}x_n^\nu(s)}{\mathrm{d}s}}\, \mathrm{d}s - \int_{\mathrm{M}} \left(\tfrac{1}{4\mu_0} \mathrm{F}_{\mu\nu}\mathrm{F}^{\mu\nu} + j_\mu A^\mu \right) \mathrm{d}^4x \ . \qquad (424)$$

What is new is the measure of the change produced by the electromagnetic field. Its internal change is given by the term F^*F, and the change due to interaction with matter is given by the term jA.

The least action principle, as usual, states that the change in a system is always as small as possible. The action S_{CED} leads to the evolution equations by requiring that the action be stationary under variations δ and δ' of the positions and of the fields which vanish at infinity. In other terms, the principle of least action requires that

$$\begin{aligned} \delta S = 0 \quad & \text{when} \quad x_\mu = x_\mu + \delta_\mu \quad \text{and} \quad A_\mu = A_\mu + \delta'_\mu \quad , \\ & \text{provided} \quad \delta x_\mu(\theta) \to 0 \quad \text{for} \quad |\theta| \to \infty \\ & \text{and} \quad \delta A_\mu(x_\nu) \to 0 \quad \text{for} \quad |x_\nu| \to \infty \ . \end{aligned} \qquad (425)$$

Page 165
Challenge 945 ny

In the same way as in the case of mechanics, using the variational method for the two variables A and x, we recover the evolution equations for particle and fields

$$b^\mu = \frac{q}{m} \mathrm{F}^\mu_\nu u^\nu \quad , \quad \partial_\mu \mathrm{F}^{\mu\nu} = j^\nu \sqrt{\frac{\mu_0}{\varepsilon_0}} \quad , \quad \text{and} \quad \varepsilon^{\mu\nu\rho\sigma} \partial_\nu \mathrm{F}_{\rho\sigma} = 0 \ , \qquad (426)$$

which we know already. Obviously, they are equivalent to the variational principle based on S_{CED}. Both descriptions have to be completed by specifying *initial conditions* for the particles and the fields, as well as *boundary conditions* for the latter. We need the first and zeroth derivatives of the position of the particles, and the zeroth derivative for the electromagnetic field.

Are you able to specify the Lagrangian of the pure electrodynamic field using the fields **E** and **B** instead of F and $^*F$?

Challenge 946 ny

The form of the Lagrangian implies that electromagnetism is *time reversible*. This means that every example of motion due to electric or magnetic causes can also take place backwards. This is easily deduced from the properties of the Lagrangian. On the other hand, everyday life shows many electric and magnetic effects which are not time invariant, such as the breaking of bodies or the burning of electric light bulbs. Can you explain how this fits together?

Challenge 947 ny

In summary, with the Lagrangian (423) all of classical electrodynamics can be described and understood. For the rest of this chapter, we look at some specific topics from this vast field.

Symmetries: the energy–momentum tensor

We know from classical mechanics that we get the definition of energy and momentum tensor by using Noether's theorem, if we determine the conserved quantity from the Lorentz symmetry of the Lagrangian. For example, we found that relativistic particles have an energy–momentum *vector*. At the point at which the particle is located, it describes the energy and momentum.

Since the electromagnetic field is not a localized entity, like a point particle, but an extended entity, we need to know the *flow* of energy and momentum at every point in space, separately *for each direction*. This makes a description with a *tensor* necessary.

$$T^{\mu\nu} = \left(\begin{array}{c|c} u & S/c = cp \\ \hline cp & T \end{array}\right) = \left(\begin{array}{c|c} 0 & \varepsilon_0 c\boldsymbol{E}\times\boldsymbol{B} \\ \hline \varepsilon_0 c\cdot & -\varepsilon_0 E_i E_j - B_i B_j/\mu_0 \\ \boldsymbol{E}\times\boldsymbol{B} & 1/2\delta_{ij}(\varepsilon_0 E^2 + B^2/\mu_0) \end{array}\right) \tag{427}$$

Both the Lagrangian and the energy momentum tensor show that electrodynamics is symmetric under motion inversion. If all charges change direction of motion – a situation often incorrectly called 'time inversion' – they move backwards.

The Lagrangian and the energy momentum tensor also show that electrodynamics is Lorentz and gauge invariant.

We also note that charges and mass destroy the symmetry of the vacuum that we mentioned in special relativity. Only the vacuum is invariant under conformal symmetries; in particular, only the vacuum is invariant under the spatial inversion $r \to 1/r$.

To sum up, electrodynamic motion, like all other examples of motion that we have encountered so far, is deterministic, conserved and reversible. This is no big thing. Nevertheless, two symmetries of electromagnetism deserve special mention.

What is a mirror?

We will study the strange properties of mirrors several times during our walk. We start with the simplest one first. Everybody can observe, by painting each of their hands in a different colour, that a mirror does *not* exchange right and left, as little as it exchanges up and down; however, a mirror does exchange right and left *handedness*. In fact, it does so by exchanging front and back.

Electrodynamics give a second answer: a mirror is a device that switches magnetic north and south poles. Can you confirm this with a diagram? Challenge 948 n

But is it always possible to distinguish left from right? This seems easy: this text is quite different from a ɓɘɿoɿɿim version, as are many other objects in our surroundings. But take a simple landscape. Are you able to say which of the two pictures of Figure 235 is the original?

Astonishingly, it is actually impossible to distinguish an original picture of nature from its mirror image if it does not contain any human traces. In other words, everyday nature is somehow left–right symmetric. This observation is so common that all candidate exceptions, from the jaw movement of ruminating cows to the helical growth of plants, such as hops, or the spiral direction of snail shells, have been extensively studied.* Can you name a few more? Page 871 Challenge 949 n

* The most famous is the position of the heart. The mechanisms leading to this disposition are still being investigated. Recent research suggests that the oriented motion of the cilia on embryos, probably in the region called the *node*, determines the right–left asymmetry. The deep origin of this asymmetry is not yet elucidated, however. Ref. 484

Most human bodies have more muscles on the right side for right-handers, such as Albert Einstein

Figure 235 Which one is the original landscape?

The left–right symmetry of nature appears because everyday nature is described by gravitation and, as we will see, by electromagnetism. Both interactions share an important property: substituting all coordinates in their equations by the negative of their values leaves the equations unchanged. This means that for any solution of these equations, i.e. for any naturally occurring system, a mirror image is a possibility that can also occur naturally. Everyday nature thus cannot distinguish between right and left. Indeed, there are right *and* left handers, people with their heart on the left *and* others with their heart on the right side, etc.

To explore further this strange aspect of nature, try the following experiment: imagine you are exchanging radio messages with a Martian; are you able to explain to him what right and left are, so that when you meet, you are sure you are talking about the same thing? Challenge 950 n

Actually, the mirror symmetry of everyday nature – also called its *parity invariance* – is so pervasive that most animals cannot distinguish left from right in a deeper sense. Ref. 485 Most animals react to mirror stimuli with mirror responses. It is hard to teach them different ways to react, and it is possible almost only for mammals. The many experiments performed in this area gave the result that animals have symmetrical nervous systems, and possibly only humans show *lateralization*, i.e. a preferred hand and different uses for the left and the right parts of the brain.

To sum up this digression, classical electrodynamics is left–right symmetric, or parity invariant. Can you show this using its Lagrangian? Challenge 951 n

A concave mirror shows an inverted image; so does a plane mirror if it is partly folded along the horizontal. What happens if this mirror is rotated around the line of sight? Challenge 952 n

Why are do metals provide good mirrors? Metals are strong absorbers of light. Any strong absorber has a metallic shine. This is true for metals, if they are thick enough, but also for dye or ink crystals. Any material that strongly absorbs a light wavelength also reflects it efficiently. The cause of the strong absorption of a metal is the electrons inside

---

and Pablo Picasso, and correspondingly on the left side for left-handers, such as Charlie Chaplin and Peter Ustinov. This asymmetry reflects an asymmetry of the human brain, called lateralization, which is essential to human nature.

Another asymmetry of the human body is the hair whirl on the back of the head; the majority of humans have only one, and in 80 % of the cases it is left turning. But many people have more than one.

it; they can move almost freely and thus absorb most visible light frequencies.

What is the difference between electric and magnetic fields?

Obviously, the standard answer is that electric fields have sources, and magnetic fields do not; moreover, magnetic fields are small relativistic effects of importance only when charge velocities are high or when electrical fields cancel out.

For situations involving matter, this clear distinction is correct. Up to the present day,no particle with a magnetic charge, called a *magnetic monopole*, has ever been found, even though its existence is possible in several unified models of nature. If found, the action (423) would have to be modified by the addition of a fourth term, namely the magnetic current density. However, no such particle has yet been detected, despite intensive search efforts.

Page 874

In empty space, when matter is not around, it is possible to take a completely different view. In empty space the electric and the magnetic fields can be seen as two faces of the same quantity, since a transformation such as

$$\begin{aligned} \mathbf{E} &\rightarrow c\,\mathbf{B} \\ \mathbf{B} &\rightarrow -\mathbf{E}/c \end{aligned} \tag{428}$$

called (electromagnetic) *duality* transformation, transforms each vacuum Maxwell equation into the other. The minus sign is necessary for this. (In fact, there are even more such transformations; can you spot them?) Alternatively, the duality transformation transforms F into *F. In other words, in empty space we *cannot* distinguish electric from magnetic fields.

Challenge 953 n

Matter would be symmetric under duality only if magnetic charges, also called magnetic monopoles, could exist. In that case the transformation (428) could be extended to

$$c\rho_{\mathrm{e}} \rightarrow \rho_{\mathrm{m}} \quad , \quad \rho_{\mathrm{m}} \rightarrow -c\rho_{\mathrm{e}} \; . \tag{429}$$

It was one of the great discoveries of theoretical physics that even though classical electrodynamics with matter is not symmetric under duality, nature is. In 1977, Claus Montonen and David Olive showed that quantum theory allows duality transformations even *with* the inclusion of matter. It has been known since the 1930s that quantum theory allows magnetic monopoles. We will discover the important ramifications of this result in the third part of the text. This duality turns out to be one of the essential stepping stones that leads to a unified description of motion. (A somewhat difficult question: extending this duality to quantum theory, can you deduce what transformation is found for the fine structure constant, and why it is so interesting?)

Challenge 954 ny

Duality, by the way, is a symmetry that works *only* in Minkowski space-time, i.e. in space-times of 3 + 1 dimensions. Mathematically, duality is closely related to the existence of quaternions, to the possibility of interpreting Lorentz boosts as rotations in 3 + 1 dimensions, and last, but not least, to the possibility of defining other smooth mathematical structures than the standard one on the space R^4. These mathematical connections are mysterious for the time being; they somehow point to the special role that *four* space-time dimensions play in nature. More details will become apparent in the third part of our mountain ascent.

Electrodynamic challenges and curiosities

Could electrodynamics be different?

Any interaction such as Coulomb's rule (390), which acts, for one given observer, between two particles independently of 3-velocity, must depend on 3-velocity for other inertial observers.* It turns out that such an interaction cannot be independent of the 4-velocity either. Such an interaction, even though it would indeed be 3-velocity dependent, would change the rest mass, since the 4-acceleration would not be 4-orthogonal to the 4-velocity.

The next simplest case is the one in which the acceleration is proportional to the 4-velocity. Together with the request that the interaction leaves the rest mass constant, we
Ref. 574 then recover electrodynamics.

In fact, the requirements of gauge symmetry and of relativity symmetry also make it impossible to modify electrodynamics. In short, it does not seem possible to have a behaviour different from $1/r^2$ for a classical interaction.

An inverse square dependence implies a vanishing mass of light and light particles,
Ref. 575 the photons. Is the mass really zero? The issue has been extensively studied. A massive photon would lead to a wavelength dependence of the speed of light in vacuum, to deviations from the inverse square 'law', to deviations from Ampère's 'law', to the existence of longitudinal electromagnetic waves and more. No evidence for these effects has ever been found. A summary of these studies shows that the photon mass is below 10^{-53} kg, or maybe 10^{-63} kg. Some arguments are not universally accepted, thus the limit varies somewhat from researcher to researcher.

A small non-vanishing mass for the photon would change electrodynamics somewhat. The inclusion of a tiny mass poses no special problems, and the corresponding
Ref. 575 Lagrangian, the so-called *Proca Lagrangian*, has already been studied, just in case.

Strictly speaking, the photon mass cannot be said to vanish. In particular, a photon with a Compton wavelength of the radius of the visible universe cannot be distinguished from one with zero mass through any experiment. This gives a mass of 10^{-69} kg for the photon. One notes that the experimental limits are still much larger. Photons with such a small mass value would not invalidate electrodynamics as we know it.

Interestingly, a non-zero mass of the photon implies the lack of magnetic monopoles, as the symmetry between electric and magnetic fields is broken. It is therefore important on the one hand to try to improve the experimental mass limit, and on the other hand to explore whether the limit due to the universe's size has any implications for this issue. This question is still open.

The toughest challenge for electrodynamics

Electrodynamics faces an experimental and theoretical issue that physicist often avoid. The process of thought is electric in nature. Physics faces two challenges in this domain.

* This can be deduced from special relativity from the reasoning of page 497 or from the formula in the footnote of page 295.

First, physicists must find ways of modelling the thought process. Second, measurement technology must be extended to allow one to measure the currents in the brain.

Even though important research has been carried out in these domains, researchers are still far from a full understanding. Research using computer tomography has shown, for example, that the distinction between the conscious and the unconscious can be measured and that it has a biological basis. Psychological concepts such as repression can be observed in actual brain scans. Modellers of the brain mechanisms must thus learn to have the courage to take some of the concepts of psychology as descriptions for actual physical processes. This approach requires one to translate psychology into physical models, an approach that is still in its infancy.

Similarly, research into magnetoencephalography devices is making steady progress. The magnetic fields produced by brain currents are as low as 10 fT, which require sensors at liquid helium temperature and a good shielding of background noise. Also the spatial resolution of these systems needs to be improved.

The whole programme would be considered complete as soon as, in a distant future, it was possible to use sensitive measuring apparatus to detect what is going on inside the brain and to deduce or 'read' the thoughts of a person from these measurements. In fact, this challenge might be the most complex of all challenges that science is facing. Clearly, the experiment will require involved and expensive machinery, so that there is no danger for a misuse of the technique. It could also be that the spatial resolution required is beyond the abilities of technology. However, the understanding and modelling of the brain will be a useful technology in other aspects of daily life as well.*

15. What is light?

The nature of light has fascinated explorers of nature at least since the time of the ancient Greeks. In 1865, Maxwell summarized all data collected in the 2500 years before him by deducing an basic consequence of the equations of electrodynamics. He found that in the case of vacuum, the equations of the electrodynamic field could be written as

Ref. 486

$$\square\ \mathbf{A} = 0 \quad \text{or, equivalently} \quad \varepsilon_0\mu_0 \frac{\partial^2\varphi}{\partial t^2} + \frac{\partial^2 A_x}{\partial x^2} + \frac{\partial^2 A_y}{\partial y^2} + \frac{\partial^2 A_z}{\partial z^2} = 0 \; . \tag{430}$$

Challenge 955 e

This is called a *wave equation*, because it admits solutions of the type

$$\mathbf{A}(t,\mathbf{x}) = \mathbf{A}_0 \sin(\omega t - \mathbf{k}\mathbf{x} + \delta) = \mathbf{A}_0 \sin(2\pi f t - 2\pi \mathbf{x}/\lambda + \delta) \; , \tag{431}$$

which are commonly called *plane waves*. Such a wave satisfies equation (430) for any value of the *amplitude* A_0, of the *phase* δ, and of the *angular frequency* ω, provided the *wave vector* $\mathbf{k}$ satisfies the relation

$$\omega(\mathbf{k}) = \frac{1}{\sqrt{\varepsilon_0\mu_0}} k \quad \text{or} \quad \omega(\mathbf{k}) = \frac{1}{\sqrt{\varepsilon_0\mu_0}} \sqrt{\mathbf{k}^2} \; . \tag{432}$$

* This vision, formulated here in 2005, is so far from realization that it is unclear whether it will come true in the twenty-first or in any subsequent century.

The relation $\omega(\mathbf{k})$ between the angular frequency and the wave vector, the so-called *dispersion relation*, is the main property of any type of wave, be it a sound wave, a water wave, an electromagnetic wave, or any other kind. Relation (432) specifically characterizes electromagnetic waves in vacuum, and distinguishes them from all other types of waves.*

Equation (430) for the electromagnetic field is *linear* in the field; this means that the sum of two situations allowed by it is itself an allowed situation. Mathematically speaking, any *superposition* of two solutions is a solution as well. For example, this means that two waves can cross each other without disturbing each other, and that waves can travel across static electromagnetic fields. Linearity also means that any electromagnetic wave can be described as a superposition of pure sine waves, each of which is described by expression (431).

Heinrich Hertz

After Maxwell predicted the existence of electromagnetic waves, in the years between 1885 and 1889 Heinrich Hertz* discovered and studied them. He fabricated a very simple transmitter and receiver for 2 GHz waves. Waves around this frequency are used in the last generation of mobile phones. These waves are now called *radio waves*, since physicists tend to call all moving force fields *radiation*, recycling somewhat incorrectly a Greek term which originally means 'light emission.'

Hertz also measured the speed of these waves. In fact, you can measure the speed also at home, with a chocolate bar and a kitchen microwave oven. A microwave oven emits radio waves at 2.5 GHz – not far from Hertz's value. Inside the oven, these waves form standing waves. Just put the chocolate bar (or a piece of cheese) in the oven and switch the power off as soon as melting begins. You will notice that the bar melts at regularly spaced spots. These spots are half a wavelength apart. From the wavelength and the frequency, the speed of light and radio waves simply follows as the product of the two.

If you are not convinced, you can measure the speed directly, by telephoning to a friend on another continent, if you can make sure to use a satellite line (just chose a low cost provider). There is about half a second additional delay between the end of a sentence and the answer of the friend, compared to normal conversation. In this half second, the signal goes up to the geostationary satellite, down again and the same way back. This half second gives a speed of $c \approx 4 \cdot 36\,000\,\mathrm{km}/0.5\,\mathrm{s} \approx 3 \cdot 10^5\,\mathrm{km/s}$, which is close to the precise value. Radio amateurs who reflect their signals on the Moon can perform more precise measurements. A different method, for computer fans, uses the 'ping' command.
Ref. 487
The 'ping' command measures the time for a computer signal to get to another computer and back. If the cable length between two computers is known, the signal speed can be
Challenge 956 e
deduced. Just try.

But Maxwell did more. He strengthened earlier predictions that *light* itself is a solution of equation (431) and therefore an electromagnetic wave, albeit with a much higher

Page 189
* For completeness, a *wave* in physics is any propagating imbalance.

* Heinrich Rudolf Hertz (b. 1857 Hamburg, d. 1894 Bonn), important Hamburger theoretical and experimental physicist. The unit of frequency is named after him. Despite his early death, Hertz was a central figure in the development of electromagnetism, in the explanation of Maxwell's theory and in the unfolding of radio communication technology. More about him on page 139.

Figure 237 The primary, secondary and supernumerary rainbows (© Wolfgang Hinz)

frequency. Let us see how we can check this.

It is easy to confirm the *wave* properties of light; indeed they were known already long before Maxwell. In fact, the first to suggest that light is a wave was, around the year 1678, the important Dutch physicist Christiaan Huygens (b. 1629 's Gravenhage, d. 1695 Hofwyck). You can confirm this fact with your own fingers. Simply put your hand one or two centimetres in front of the eye, look towards the sky through the gap between the middle finger and the index and let the two fingers almost touch. You will see a number of dark lines dividing the gap. These lines are the interference pattern formed by the light behind the slit created by the fingers. *Interference* is the name given to those amplitude patterns which appear when several waves superpose.* The interference patterns depends on the spacing between the fingers. This experiment therefore allows to estimate the wavelength of light, and thus, if you know its speed, also its frequency.
Challenge 958 n Are you able to do so?

Figure 236 The first transmitter (left) and receiver (right) of electromagnetic (micro-) waves

Historically, another effect was central in convincing everybody that light was a wave: the supernumerary rainbows, the additional bows below the main or primary rainbow. If we look carefully at a rainbow, below the main red–yellow–green–blue–violet bow, we
Ref. 488 observe weaker, additional green blue and violet bows. Depending on the intensity of the rainbow, several of these supernumerary rainbows can be observed. They are due to an interference effect, as Thomas Young showed around 1803.** Indeed, the repetition distance of the supernumerary bows depends on the radius of the average water droplets

Challenge 957 n * Where does the energy go in interference patterns?

** Thomas Young (1773 Milverton–1829), read the bible at two, spoke Latin at four; doctor of medicine, he

Page 531 that form them. (Details about the normal rainbows are given below.) Supernumerary rainbows were central in convincing people that light is a wave. It seems that in those times scientists either did not trust their own fingers, or did not have any.

There are many other ways that the wave character of light becomes apparent. Maybe the most beautiful is an experiment carried out by a team of Dutch physicists in 1990. They simply measured the light transmitted through a *slit* in a metal plate. It turns out that the transmitted intensity depends on the width of the slit. Their surprising result is shown in Figure 238. Can you explain the origin of the unexpected intensity steps in the curve? Ref. 489 Challenge 959 ny

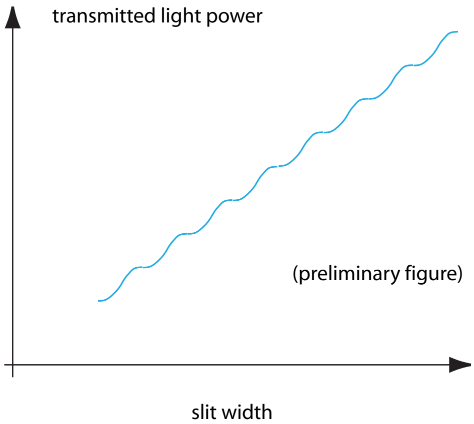


Figure 238 The light power transmitted through a slit as function of its width

Numerous other experiments on the creation, detection and measurement of electromagnetic waves have been performed in the nineteenth and twentieth century. For example, in 1800, William Herschel discovered *infrared light* using a prism and a thermometer. (Can you guess how?) In 1801, Johann Wilhelm Ritter (1776–1810) a colourful figure of natural Romanticism, discovered *ultraviolet light* using silver chloride, AgCl, and again a prism. The result of all these experiments is that electromagnetic waves must be distinguished above all by their wavelength or frequency. The main categories are listed in Table 46. For visible light, the wavelength lies between 0.4 μm (pure violet) and 0.8 μm (pure red). Challenge 960 n Page 521

At the end of the twentieth century the final confirmation of the wave character of light become possible. Using quite sophisticated experiments researchers measured the oscillation frequency of light *directly*. The value, between 375 and 750 THz, is so high that detection was impossible for a long time. But with these modern experiments the dispersion relation (432) of light has finally been confirmed in all its details. Ref. 490

We are left with one additional question about light. If light oscillates, in which direction does this happen? The answer is hidden in the parameter $\mathbf{A}_0$ in expression (431). Electromagnetic waves oscillate in directions *perpendicular* to their motion. Therefore, even for identical frequency and phase, waves can still differ: they can have different *polarization* directions. For example, the polarization of radio transmitters determines whether radio antennas of receivers have to be kept horizontal or vertical. Also for light, polarization is easily achieved, e.g. by shining it through stretched plastic films. When the polarization of light was discovered in 1808 by the French physicist Louis Malus (1775–

became professor of physics. He introduced the concept of *interference* into optics, explaining the Newtonian rings and the supernumerary rainbows; he was the first person to determine light's *wavelength*, a concept that he also introduced, and its dependence on colour. He was the first to deduce the three colour vision explanation of the eye and after reading of the discovery of polarization, explained light as a transverse wave. In short, Young discovered most what people learn at school about light. He was a universal talent: he also worked on the deciphering of hieroglyphs, on ship building and on engineering problems. In Britain his ideas on light were not accepted, since Newton and his followers crushed all opposing views. Young collaborated with Fraunhofer and Fresnel; at last, his results were made known by Fresnel and Helmholtz.

1812), it definitively established its wave nature. Malus discovered it when he looked at the strange double images produced by feldspar, a transparent crystal found in many minerals. Feldspar ($KAlSi_3O_8$) splits light beams into two – it is *birefringent* – and polarizes them differently. That is the reason that feldspar is part of every crystal collection. Calcite ($CaCO_3$) shows the same effect. If you ever see a piece of feldspar or transparent calcite, have a look through it onto some written text.

By the way, the human eye is unable to detect polarization, in contrast to many insects, spiders and certain birds. Honey bees use polarization to deduce the position of the Sun even when it is hidden behind clouds, some beetles of the genus *Scarabeus* use the polarization of the Moon light for navigation, and many insects use polarization to distinguish water surfaces from mirages. Can you find out how? Despite the human inability to detect polarization, both the cornea and the lens of the human eye are birefringent.

Ref. 491
Challenge 961 n
Ref. 492

Note that all possible polarizations of light form a continuous set. However, a general plane wave can be seen as a superposition of two orthogonal, linearly polarized waves with different amplitudes and different phases. Most books show pictures of plane, linearized electrodynamic waves. Essentially, the electric fields look like water waves generalized to three dimensions, the same for magnetic fields, and the two are perpendicular to each other. Can you confirm this?

Challenge 962 ny

Interestingly, a generally polarized plane wave can also be seen as the superposition of right and left *circularly polarized waves*. However, no figures of such waves are found in any textbook. Can you explain why?

Challenge 963 n

So far it is clear that light is a wave. To confirm that light waves are indeed *electromagnetic* is more difficult. The first argument was given by Bernhard Riemann in 1858;* he deduced that any electromagnetic wave must propagate with a speed

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} . \tag{433}$$

Already ten years before him, in 1848, Kirchoff had noted that the measured values on both sides agreed within measurement errors. A few years later, Maxwell gave a beautiful confirmation by deducing the expression from equation (432). You should be able to repeat the feat. Note that the right hand side contains electric and magnetic quantities, and the left hand side is an optical entity. Riemann's expression thus unifies electromagnetism with optics.

Challenge 964 ny

Of course, people were not yet completely convinced. They looked for more ways to show that light is electromagnetic in nature. Now, since the evolution equations of the electrodynamic field are linear, additional electric or magnetic fields alone do not influence the motion of light. On the other hand, we know that electromagnetic waves are emitted only by accelerated charges, and that all light is emitted from matter. We thus follow that matter is full of electromagnetic fields and accelerated electric charges. This in turn implies that the influence of matter on light can be understood from its internal electromagnetic fields, and in particular, that subjecting matter to *external* electromagnetic

Challenge 965 e

* Bernhard Riemann (b. 1826 Breselenz, d. 1866 Selasca), important German mathematician. He studied curved space, providing several of the mathematical and conceptual foundations of general relativity, but then died at an early age.

fields should change the light it emits, the way matter interacts with light, or generally, the material properties as a whole.

Searching for effects of electricity and magnetism on matter has been a main effort of physicists for over hundred years. For example, electric fields influence the light transmission of oil, an effect discovered by John Kerr in 1875.* The discovery that certain gases change colour when subject to a field yielded several Nobel prizes for physics. With time, many more influences on light related properties by matter subjected to fields were found. An extensive list is given in the table on page 558. It turns out that apart from a few exceptions the effects can *all* be described by the electromagnetic Lagrangian (423), or equivalently, by Maxwell's equations (426). In summary, classical electrodynamics indeed unifies the description of electricity, magnetism and optics; all phenomena in these fields, from the rainbow to radio and from lightning to electric motors, are found to be different aspects of the evolution of the electromagnetic field.

Table 46 The electromagnetic spectrum

Frequency	Wavelength	Name	Main properties	Appearance	Use
$3 \cdot 10^{-18}$ Hz	10^{26} m	**lower frequency limit**		see section on cosmology	
< 10 Hz	> 30 Mm	**quasistatic fields**		intergalactic, galactic, stellar and planetary fields, brain, electrical fish	power transmission, accelerating and deflecting cosmic radiation
		radio waves		electronic devices	
10 Hz-50 kHz	30 Mm-6 km	ELW	go round the globe, penetrate into water, penetrate metal	nerve cells, electromechanical devices	power transmission, communication through metal walls, communication with submarines http://www.vlf.it
50 - 500 kHz	6 km-0.6 km	LW	follow Earth curvature, felt by nerves ('bad weather nerves')	emitted by thunderstorms	radio communications, telegraphy, inductive heating
500 - 1500 kHz	600 m-200 m	MW	reflected by night sky		radio
1.5 - 30 MHz	200 m-10 m	SW	circle world if reflected by the ionosphere, destroy hot air balloons	emitted by stars	radio transmissions, radio amateurs, spying

* John Kerr (1824–1907), Scottish physicist, friend and collaborator of William Thomson.

FRE-QUENCY	WAVE-LENGTH	NAME	MAIN PROPERTIES	APPEARANCE	USE
15 - 150 MHz	20 m-2 m	VHF	allow battery operated transmitters	emitted by Jupiter	remote controls, closed networks, tv, radio amateurs, radio navigation, military, police, taxi
150 - 1500 MHz	2 m-0.2 m	UHF	idem, line of sight propagation		radio, walkie-talkies, tv, mobile phones, internet via cable, satellite communication, bicycle speedometers
		microwaves			
1.5 - 15 GHz	20 cm-2 cm	SHF	idem, absorbed by water	night sky, emitted by hydrogen atoms	radio astronomy, used for cooking (2.45 GHz), telecommunications, radar
15 - 150 GHz	20 mm-2 mm	EHF	idem, absorbed by water		
		infrared	go through clouds	emitted by every warm object	satellite photography of Earth, astronomy
0.3 - 100 THz	1000 -3 μm	IRC or far infrared		sunlight, living beings	seeing through clothes, envelopes and teeth
100 - 210 THz	3 μm-1.4 μm	IRB or medium infrared		sunlight	used for optical fibre communications for telephone and cable TV
210 - 385 THz	1400-780 nm	IRA or near infrared	penetrates for several cm into human skin	sunlight, radiation from hot bodies	healing of wounds, rheumatism, sport physiotherapy, hidden illumination
375 - 750 THz	800-400 nm	**light**	not absorbed by air, detected by the eye (up to 850 nm at sufficient power)	heat (‘hot light’), lasers & chemical reactions e.g. phosphor oxidation, fireflies (‘cold light’)	definition of straightness, enhancing photosynthesis in agriculture, photodynamic therapy, hyperbilirubinaemia treatment
375 - 478 THz	780-627 nm	red	penetrate flesh	blood	alarm signal, used for breast imaging

Frequency	Wavelength	Name	Main properties	Appearance	Use
	700 nm	pure red		rainbow	colour reference for printing, painting, illumination and displays
478 - 509 THz	627-589 nm	orange		various fruit	attracts birds and insects
	600 nm	standard orange			
509 - 530 THz	589-566 nm	yellow		majority of flowers	idem; best background for reading black text
	580 nm	standard yellow			
530 - 606 THz	566-495 nm	green	maximum eye sensitivity	algae and plants	highest brightness per light energy for the human eye
	546.1 nm	pure green		rainbow	colour reference
606 - 688 THz	495-436 nm	blue		sky, gems, water	
	488 nm	standard cyan			
	435.8 nm	pure blue		rainbow	colour reference
688 - 789 THz	436-380 nm	indigo, violet		flowers, gems	
		ultraviolet			
789 - 952 THz	380-315 nm	UVA	penetrate 1 mm into skin, darken it, produce vitamin D, suppress immune system, cause skin cancer, destroy eye lens	emitted by Sun and stars	seen by certain birds, integrated circuit fabrication
0.95 - 1.07 PHz	315-280 nm	UVB	idem, destroy DNA, cause skin cancer	idem	idem
1.07 - 3.0 PHz	280-100 nm	UVC	form oxygen radicals from air, kill bacteria, penetrate 10 μm into skin	idem	disinfection, water purification, waste disposal, integrated circuit fabrication
3 -24 PHz	100-13 nm	EUV			sky maps, silicon lithography
		X-rays	penetrate materials	emitted by stars, plasmas and black holes	imaging human tissue

Frequency	Wavelength	Name	Main properties	Appearance	Use
24 - 240 PHz	13-1.3 nm	soft X-rays	idem	synchrotron radiation	idem
> 240 PHz or > 1 keV	< 1.2 nm	hard X-rays	idem	emitted when fast electrons hit matter	crystallography, structure determination
> 12 EHz or > 50 keV	< 24 pm	γ-rays	idem	radioactivity, cosmic rays	chemical analysis, disinfection, astronomy
$2 \cdot 10^{43}$ Hz	$\approx 10^{-35}$ m	**Planck limit**		see part three of this text	

The slowness of progress in physics

The well-known expression

$$c = 1/\sqrt{\varepsilon_0 \mu_0} \qquad (434)$$

for the speed of light is so strange that one should be astonished when one sees it. Something essential is missing.

Indeed, the speed is *independent* of the proper motion of the observer measuring the electromagnetic field. In other words, the speed of light is independent of the speed of the lamp and independent of the speed of the observer. All this i contained in expression (434). Incredibly, for five decades, *nobody* explored this strange result. In this way, the theory of relativity remained undiscovered from 1848 to 1905. Like in so many other cases, the progress of physics was much slower than necessary.

At the end of the nineteenth century, the teenager Albert Einstein, teenager, read a book discussing the issue of the constancy of the speed of light. Like the book, Einstein asked himself what would happen if an observer would move at the same speed as light itself, and in particular, what kind of electromagnetic field he would observe in that case. Einstein later explained that this Gedanken experiment convinced him already at that young age that *nothing* could travel at the speed of light, since the field observed would have a property not found in nature. Can you find out which one he meant? Challenge 966 n

The constancy of the speed of light is the essential point that distinguishes special relativity from Galilean physics. In this sense, any electromagnetic device, making use of expression (434), is a working proof of special relativity.

Does light travel in a straight line?

Usually light moves in straight lines. Indeed, we even use light to *define* 'straightness.' However, there is a number of exceptions which every expert on motion should know. Ref. 493

In sugar syrup, light beams curve, as shown in Figure 239. In fact, light beams bend at any material interface. This effect, called *refraction*, also changes the shape of our feet when we are in the bath tub and makes aquaria seem less deep than they actually are. Refraction is a consequence of the change of light speed from material to material. Are

Figure 240 Virtual and real images

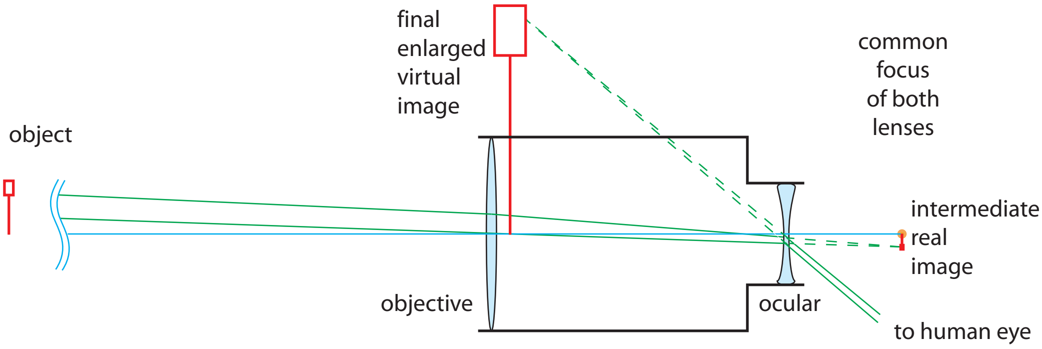


Figure 241 Refraction as the basis of the telescope – shown here in the original Dutch design

Challenge 967 n

you able to explain refraction, and thus explain the syrup effect?

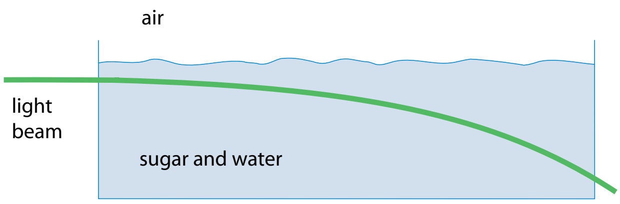


Figure 239 Sugar water bends light

Refraction is chiefly used in the design of lenses. Using glass instead of water, one can produce curved surfaces, so that light can be *focussed*. Focussing devices can be used to produce images. The two main types of lenses, with their focal points and the images they produce, are shown in Figure 240. When an object is put between a converging lens and its focus, the lens produces a real image and works as a *magnifying glass*. It also produces a *real* image, i.e., an image that can be projected on a screen. In all other cases lenses produce so-called *virtual images*: such images can be seen with the eye but not be projected onto a screen.

Even though glasses and lenses were known since antiquity, the middle ages had to pass by before two lenses were combined to make more elaborate optical instruments. The *telescope* was invented in or just before 1608 by the German–Dutch lens grinder Johannes Lipperhey (*c.* 1570–1619), who made a fortune by selling it to the Dutch military. When Galileo heard about the discovery, he quickly took it over and improved it. Already in 1609 he made the first astronomical observations which rendered him world-famous. The Dutch design yields a short tube and a bright and upright image. It is still used today in opera glasses. Many other ways to build telescopes have been developed over the years.*

* A fascinating overview about what people do in this domain up to this day are told by P. Manly, Unusual

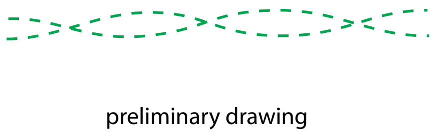


Figure 242 In certain materials, light beams can spiral around each other

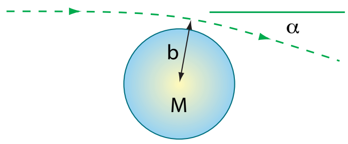


Figure 243 Masses bend light

Another way to combine two lenses leads to the *microscope*. Can you explain to a non-physicist how a microscope works?* Werner Heisenberg almost missed his Ph.D. exam because he could not. The problem is not difficult, though. Indeed, the inventor of the microscope was an autodidact of the seventeenth century: the Dutch technician Antoni van Leeuwenhoek (1632–1723) made a living by selling over five hundred of his microscopes to his contemporaries.

Challenge 968 n

Refraction is often colour-dependent. For that reason, microscopes or photographic cameras have several lenses, made of different materials. They compensate the colour effects, which otherwise yield coloured image borders. The colour dependence of refraction in water droplets is also the basis of the rainbow, as shown below, and refraction in ice crystals in the atmosphere is at the basis of the halos and the many other light patterns often seen around the Sun and the Moon.

Page 531

Ref. 494

A second important observation is that light goes around corners, and the more so the more they are sharp. This effect is called *diffraction*. In fact, light goes around corners in the same way that sound does. Diffraction is due to the wave nature of light (and sound). You probably remember the topic from school.

Because of diffraction, it is impossible to produce strictly parallel light beams. For example, every laser beam diverges by a certain minimum amount, called the *diffraction limit*. Maybe you know that the world's most expensive Cat's-eyes are on the Moon, where they have been deposited by the Apollo 11 cosmonauts and the Lunakhod mission. Can you determine how wide a laser beam with minimum divergence has become when it arrives at the Moon and back on Earth, assuming that it was 1 m wide when departing from Earth? How wide would it come back if it had been 1 mm wide at the start?

Ref. 495

Challenge 969 n

Diffraction implies that there are no perfectly sharp images: there exists a *limit on resolution*. This is true for every optical instrument, including the eye. The resolution of the eye is between one and two minutes of arc, i.e. between 0.3 and 0.6 mrad. The limit is due to the finite size of the pupil. Therefore for example, there is a maximum distance at which humans can distinguish the two headlights of a car. Can you estimate it?

Challenge 970 n

Resolution limits also make it impossible to see the Great Wall in northern China from the Moon, contrary to what is often claimed. In the few parts which are not yet in ruins,

telescopes, Cambridge University Press, 1991. Images can also be made with mirrors. Since mirrors are cheaper and more easy to fabricate with high precision, most large telescopes have a mirror instead of the first lens.

By the way, telescopes also exist in nature. Many spiders have two types of eyes. The large ones, made to see far away, have two lenses arranged in the same way as in the telescope.

* If not, read the beautiful text by Elizabeth M. Slater & Henry S. Slater, *Light and Electron Microscopy*, Cambridge University Press, 1993.

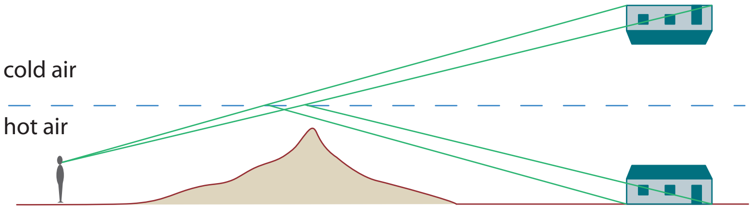


Figure 244 Reflection at air interfaces is the basis of the Fata Morgana

the wall is about 6 metres wide, and even if it casts a wide shadow during the morning or the evening, the angle it subtends is way below a second of arc, so that it is completely invisible to the human eye. In fact, three different cosmonauts who went to the Moon performed careful searches and confirmed that the claim is absurd. The story is one of the most tenacious urban legends. (Is it possible to see the wall from the space shuttle?) The largest man-made objects are the polders of reclaimed land in the Netherlands; they *are* visible from outer space. So are most large cities as well as the highways in Belgium at night; their bright illumination makes them stand out clearly from the dark side of the Earth.

Ref. 496
Challenge 971 ny

Diffraction also means that behind a small disk illuminated along its axis, the centre of the shadow shows, against all expectations, a bright spot. This 'hole' in the shadow was predicted in 1819 by Denis Poisson (1781–1840) in order to show to what absurd consequences the wave theory of light would lead. He had just read the mathematical description of diffraction developed by Augustin Fresnel* on the basis of the wave description of light. But shortly afterwards, François Arago (1786–1853) actually observed Poisson's point, converting Poisson, making Fresnel famous and starting the general acceptance of the wave properties of light.

Additional electromagnetic fields usually do not influence light directly, since light has no charge and since Maxwell's equations are linear. But in some materials the equations are non-linear, and the story changes. For example, in certain photorefractive materials, two nearby light beams can even *twist* around each other, as shown by Segev and coworkers in 1997.

Ref. 497

A final way to bend light is gravity, as discussed already in the chapters on universal gravity and on general relativity. Also the effect of gravity between two light beams was discussed there.

Page 440

In summary, light travels straight only if it travels *far from other matter*. In everyday life, 'far' simply means more than a few millimetres, because all gravitational and electromagnetic effects are negligible at these distances, mainly due to lights' truly supersonic speed.

The concentration of light

* Augustin Jean Fresnel (1788–1827), engineer and part time physicist; he published in 1818 his great paper on wave theory for which he got the prize of the French academy of sciences in 1819. To improve his finances, he worked in the commission responsible for lighthouses, for which he developed the well-known Fresnel lens. He died prematurely, partly of exhaustion due to overwork.

If one builds a large lens or a curved mirror, one can collect the light of the Sun and focus it on a single spot. Everybody has used a converging lens as a child to burn the black spots of newspapers in this way. In Spain, people with more money have even built a curved mirror as large as a house, in order to study solar energy use and material behaviour at high temperature. Essentially, the mirror provides a cheap way to fire an oven. Indeed, 'focus' is the Latin word for 'oven'.

Figure 245 The last mirror of the solar furnace at Odeillo, in the French Pyrenees (© Gerhard Weinrebe)

Kids find out quite rapidly that large lenses allow to burn things more easily than small ones. It is obvious that the Spanish site is the record holder in this game. However, building a larger mirror does not make sense. Whatever its size may be, such a set-up cannot reach a larger temperature than that of the original light source. The surface temperature of the Sun is about 5800 K; indeed, the highest temperature reached so far is about 4000 K. Are you able to show that this limitation follows from the second law of thermodynamics?

Challenge 972 ny

In short, nature provides a *limit* to the concentration of light energy. In fact, we will encounter additional limits in the course of our exploration.

Can one touch light?

If a little glass bead is put on top of a powerful laser, the bead remains suspended in mid air, as shown in Figure 246.* That means that light has momentum. Therefore, contrary to what we said in the beginning of our mountain ascent, images *can* be touched! In fact, the ease with which objects can be pushed has even a special name. For stars, it is called the *albedo*, and for general objects it is called the *reflectivity* r.

Like each type of electromagnetic field, and like every kind of wave, light carries energy; the energy flow T per surface and time is

Challenge 973 e

$$\mathbf{T} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} \quad \text{giving an average} \quad \langle T\rangle = \frac{1}{2\mu_0}E_{\max}B_{\max} \,. \tag{435}$$

Obviously, light also has a momentum P. It is related to the energy E by

$$P = \frac{E}{c} \,. \tag{436}$$

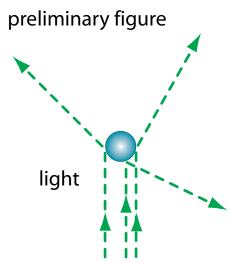


Figure 246 Levitating a small glass bead with a laser

Challenge 974 e

As a result, the pressure p exerted by light onto a body is given by

$$p = \frac{T}{c}(1+r) \tag{437}$$

* The heaviest object that has been levitated with a laser had a mass of 20 g; the laser used was enormous, and the method also made use of a few additional effects, such as shock waves, to keep the object in the air.

where for black bodies we have that a reflectivity $r = 0$ and for mirrors $r = 1$; other bodies have values in between. What is your guess for the amount of pressure due to sunlight on a black surface of one square metre? Is that the reason that we feel more pressure during the day than during the night?
Challenge 975 n

In fact, rather delicate equipment is needed to detect the momentum of light, in other words, its radiation pressure. Already around 1610, Johannes Kepler had suggested in *De cometis* that the tails of comets exist only because the light of the Sun hits the small dust particles that detach from it. For that reason, the tail always points *away* from the Sun, as you might want to check at the next opportunity. Today, we know that Kepler was right; but proving the hypothesis is not easy.
Challenge 976 e

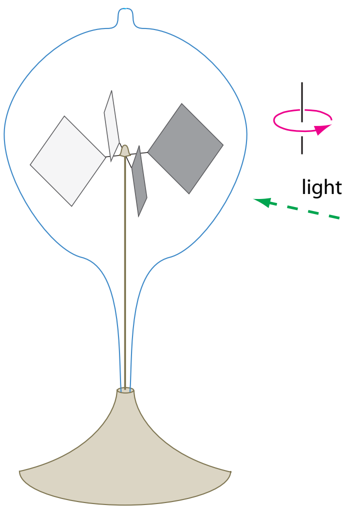


Figure 247 A commercial light mill turns *against* the light

In 1873, William Crookes * invented the *light mill radiometer*. He had the intention to demonstrate the radiation pressure of light. The light mill consists of four thin plates, black on one side and shiny on the other, which are mounted on a vertical axis, as shown in Figure 247. However, when Crookes finished building it – it was similar to those sold in shops today – he found, like everybody else, that it turned in the wrong direction, namely with the shiny side towards the light! (Why is it wrong?) You can check it by yourself by shining a laser pointer onto it. The behaviour has been a puzzle for quite some time. Explaining it involves the tiny amount of gas left over in the glass bulb and takes us too far from the topic of our mountain ascent. Only in 1901, with the advent of much better pumps, the Russian physicist Peter/Pyotr Lebedev managed to create a sufficiently good vacuum that allowed to measure the light pressure with such an improved, true radiometer. Lebedev also confirmed the predicted value of the light pressure and proved the correctness of Kepler's hypothesis. Today it is even possible to build tiny propellers that start to turn when light shines onto them, in exactly the same way that the wind turns windmills.
Challenge 977 n Ref. 498 Ref. 499 Ref. 500

But light cannot only touch and be touched, it can also *grab*. In the 1980s, Arthur Ashkin and his research group developed actual *optical tweezers* which allow to grab, to suspend and to move small transparent spheres of 1 to 20 μm diameter using laser beams. It is possible to do this through a microscope, so that one can also observe at the same time what is happening. This technique is now routinely used in biological research around the world, and has been used for example to measure the force of single muscle fibres, by chemically attaching their ends to glass or teflon spheres and then pulling them apart with such optical tweezers.
Ref. 501

But that is not all. In the last decade of the twentieth century, several groups even managed to *rotate* objects, thus realizing actual *optical spanners*. They are able to rotate
Ref. 501

* William Crookes (b. 1832 London, d. 1919 London), English chemist and physicist, president of the Royal Society, discoverer of Thallium.

particles at will in one direction or the other, by changing the optical properties of the laser beam used to trap the particle.

In fact, it does not take much to deduce that if light has linear momentum, circularly polarized light also has *angular* momentum. In fact, for such a wave the angular momentum L is given by

$$L = \frac{E_{\text{nergy}}}{\omega} . \qquad (438)$$

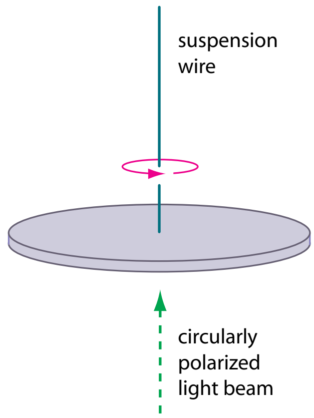


Figure 248 Light can rotate objects

Equivalently, the angular momentum of a wave is $\lambda/2\pi$ times its linear momentum. For light, this result has been confirmed already in the early twentieth century: a light beam can put certain materials (which ones?) into rotation, as shown in Figure 248. Of course, the whole thing works even better with a laser beam. In the 1960s, a beautiful demonstration was performed with microwaves. A circularly polarized microwave beam from a maser – the microwave equivalent of a laser – can put a metal piece absorbing it into rotation. Indeed, for a beam with cylindrical symmetry, depending on the sense of rotation, the angular momentum is either parallel or antiparallel to the direction of propagation. All these experiments confirm that light also carries angular momentum, an effect which will play an important role in the second part of our mountain ascent.

Challenge 978 e
Ref. 502
Challenge 979 ny

We note that not for all waves angular momentum is energy per angular frequency. This is only the case for waves made of what in quantum theory will be called spin 1 particles. For example, for gravity waves the angular momentum is *twice* this value, and they are therefore expected to be made of spin 2 particles.

In summary, light can touch and be touched. Obviously, if light can rotate bodies, it can also *be* rotated. Could you imagine how this can be achieved?

Challenge 980 n

War, light and lies

From the tiny effects of the equation (437) for light pressure we deduce that light is not an efficient tool for hitting objects. On the other hand, light is able to heat up objects, as we can feel on the skin if it is touched by a laser beam of about 100 mW or more. For the same reason even cheap laser pointers are dangerous to the eye.

In the 1980s, and again in 2001, a group of people who read too many science fiction novels managed to persuade the military – who also indulge in this habit – that lasers could be used to shoot down missiles, and that a lot of tax money should be spent to develop such lasers. Using the definition of the Poynting vector and a hitting time of about 0.1 s, are you able to estimate the weight and size of the battery necessary for such a device to work? What would happen in cloudy or rainy weather?

Challenge 981 ny

Other people tried to persuade NASA to study the possibility to propel a rocket using emitted light instead of ejected gas. Are you able to estimate whether this is feasible?

Challenge 982 ny

What is colour?

We saw that radio waves of certain frequencies are visible. Within that range, different frequencies correspond to different colours. (Are you able to convince a friend about this?) Challenge 983 n But the story is not finished here. Numerous colours can be produced either by a single wavelength, i.e. by *monochromatic* light, or by a *mixture* of several different colours. For example, standard yellow can be, if it is pure, a electromagnetic beam of 600 nm wavelength or it can be a mixture of standard green of 546.1 nm and standard red of 700 nm. The eye cannot distinguish between the two cases; only spectrometers can. In everyday life, all colours turn out to be mixed, with the exception of those of yellow street lamps, of laser beams and of the rainbow. You can check this yourself, using an umbrella or a compact disk: they decompose light mixtures, but not pure colours. Challenge 984 e

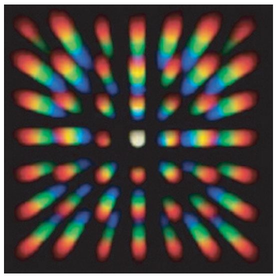


Figure 249 Umbrellas decompose white light

In particular, white light is a mixture of a continuous range of colours with a given intensity per wavelength. To check that white light is a mixture of colours, simply hold the left side of Figure 250 so near to your eye that you cannot focus the stripes any more. The unsharp borders of the white stripes have a pink or a green shade. These colours are due to the imperfections of the lens in the human eye, its so-called *chromatic aberrations*. Aberrations have the consequence that not all light frequencies follow the same path in the lens of the eye, and therefore that they hit the retina at different spots. This is the same effect that occurs in prisms or in water drops showing a rainbow. By the way, the shape of the rainbow tells something about the shape of the water droplets. Can you deduce the connection? Challenge 985 n

The right side of Figure 250 explains how rainbows form. The main idea is that internal reflection inside the water droplets in the sky are responsible for throwing back the light coming from the Sun, whereas the wavelength-dependent refraction at the air–water surface is responsible for the different paths of each colour. The first person to check this explanation was Theodoricus Teutonicus de Vriberg (*c.* 1250 to *c.* 1318), in the years from 1304 to 1310. Ref. 503 To check the explanation, he did something smart and simple; everybody can repeat this at home. Challenge 986 e He built an enlarged water droplet by filling a thin spherical glass container with water; then he shone a beam of white light through it. Theodoricus found exactly what is shown in the figure. With this experiment, he was able to reproduce the angle of the main or primary rainbow, its colour sequence, as well as the existence of a secondary rainbow, its observed angle and its inverted colour sequence.* All these bows are visible in Figure 237. Page 518 Theodoricus's beautiful experiment is sometimes called the most important contribution of natural science in the middle ages.

Even pure air splits white light. This is the reason that the sky and far away mountains are blue or that the Sun is red at sunset and at dawn. (The sky is black even during the day

Challenge 987 ny * Can you guess where the tertiary and quaternary rainbows are to be seen? There are rare reported sightings of them. The hunt to observe the fifth-order rainbow is still open. (In the laboratory, bows around droplets up to the 13th order have been observed.) Ref. 504 For more details, see the beautiful website at http://www.sundog.clara.co.uk/atoptics/phenom.htm. There are several formulas for the angles of the various orders of rainbows; they follow from geometric considerations, but are too involved to be given here.

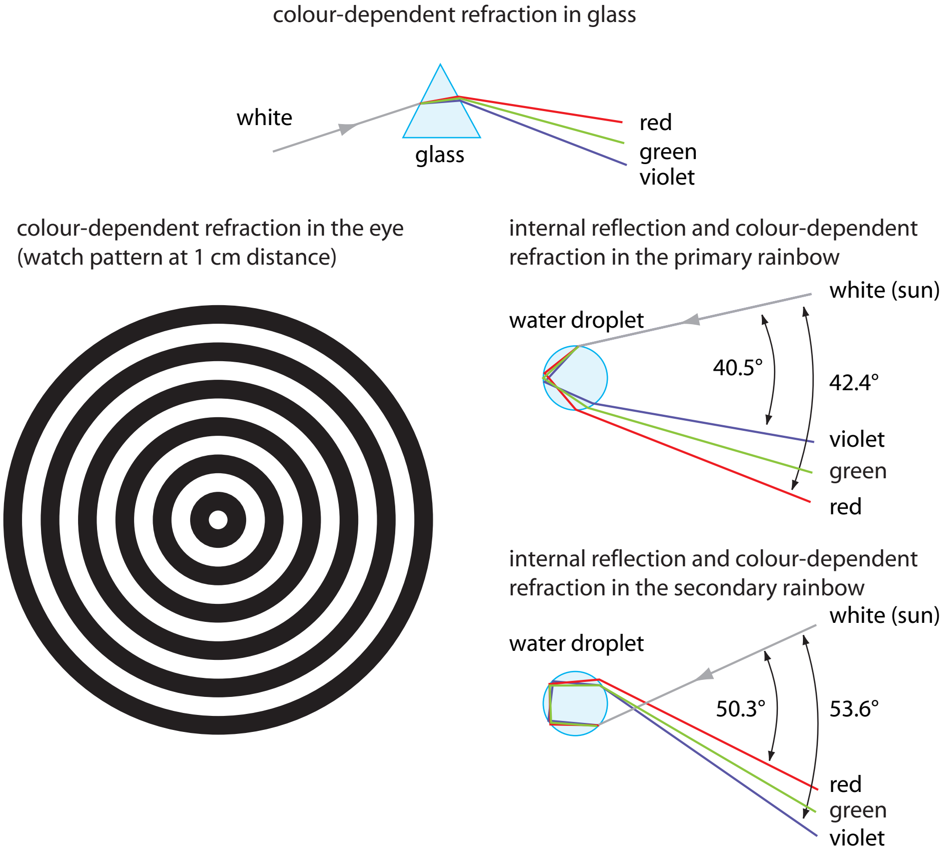


Figure 250 Proving that white light is a mixture of colours

on the Moon.) You can repeat this effect by looking through water at a black surface or at a lamp. Adding a few drops of milk to the water makes the lamp yellow and then red, and makes the black surface blue (like the sky seen from the Earth as compared to the sky seen from the Moon). More milk increases the effect. For the same reason, sunsets are especially red after volcanic eruptions.

Challenge 988 e

By the way, at sunset the atmosphere itself acts as a prism as well; that means that the Sun is split into different images, one for each colour, which are slightly shifted against each other, a bit like a giant rainbow in which not only the rim, but the whole disk is coloured. The total shift is about 1/60th of the diameter. If the weather is favourable and if the air is clear and quiet up to and beyond the horizon, for a few seconds it is possible to see, after the red, orange and yellow images of the Sun have set, the rim of the green–blue image of the Sun. That is the famous 'rayon vert' described by Jules Verne in his novel of the same title. It is often seen on islands, for example in Hawaii.*

Ref. 505

To clarify the difference between colours in physics and colour in human perception and language, a famous linguistic discovery deserves to be mentioned: colours in human

* About this and many other topics on colours in nature, such as e.g. the colour of shadows, the halos around the Moon and the Sun, and many others, see the beautiful book by Marcel Minnaert mentioned on page 71.

language have a natural *order*. Colours are ordered by all people in the world, whether from the sea, the desert or the mountains, in the following order: 1st black and white, 2nd red, 3rd green and yellow, 4th blue, 5th brown; 6th come mauve, pink, orange, grey and sometimes a twelfth term different from language to language. (Colours which point to objects, such as aubergine or sepia, or colours which are not generally applicable, such as blond, are excluded in this discussion.) The precise discovery is the following: if a particular language has a word for any of these colours, then it also has a word for all the preceding ones. The result also implies that people use these basic colour classes even if their language does *not* have a word for each of them. These strong statements have been
Ref. 506 confirmed for over 100 languages.

What is the speed of light? – Again

Physics is talking about motion. Talking is the exchange of sound; and sound is an example of a signal. A *(physical) signal* is the transport of information using transport of energy. There are no signals without motion of energy. Indeed, there is no way to store information without storing energy. To any signal we can thus ascribe a propagation speed. The highest possible signal speed is also the maximal velocity of general influences, or, using sloppy language, the maximal velocity with which effects spread causes.


Figure 251 Milk and water simulate the evening sky

If the signal is carried by matter, such as by the written text in a letter, the signal velocity is then the velocity of the material carrier, and experiments show that it is limited by the speed of light.

For a wave carrier, such as water waves, sound, light or radio waves, the situation is less evident. What is the speed of a wave? The first answer that comes to mind is the speed with which wave crests of a sine wave move. This already introduced *phase velocity* is given by the ratio between the frequency and the wavelength of a monochromatic wave, i.e. by

$$v_{ph} = \frac{\omega}{k} \ . \qquad (439)$$

For example, the phase velocity determines interference phenomena. Light in vacuum has the same phase velocity $v_{ph} = c$ for all frequencies. Are you able to imagine an ex-
Challenge 989 n periment to test this to high precision?

On the other hand, there are cases where the phase velocity is larger than c, most notably when light travels through an absorbing substance, and when at the same time the frequency is near to an absorption maximum. In these cases, experiments show that the phase velocity is *not* the signal
Ref. 507 velocity. For such situations, a better approximation to the

signal speed is the *group velocity*, i.e. the velocity at which a group maximum will travel. This velocity is given by

$$v_{gr} = \left.\frac{d\omega}{dk}\right|_{k_0} \qquad (440)$$

where k_0 is the central wavelength of the wave packet. We observe that $\omega = c(k)k = 2\pi v_{ph}/\lambda$ implies the relation

$$v_{gr} = \left.\frac{d\omega}{dk}\right|_{k_0} = v_{ph} - \lambda\frac{dv_{ph}}{d\lambda} . \qquad (441)$$

This means that the sign of the last term determines whether the group velocity is larger or smaller than the phase velocity. For a travelling group, as shown by the dotted line in Figure 252, this means that new maxima either appear at the end or at the front of the group. Experiments show that for light *in vacuum*, the group velocity has the same value $v_{gr} = c$ for all values of the wave vector k.

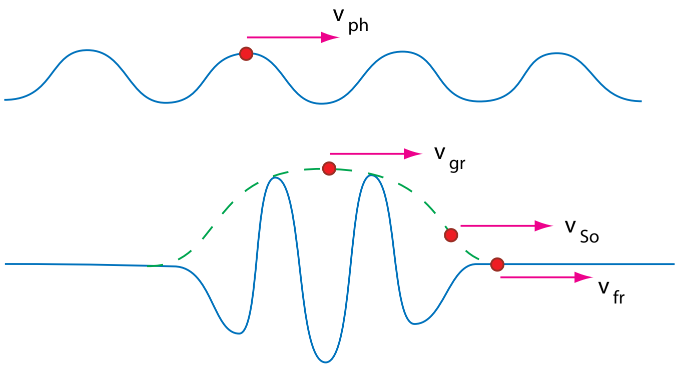


Figure 252 The definition of important velocities in wave phenomena

You should be warned that still many publications propagate the false statement that the group velocity *in materials* is never larger than c, the speed of light in vacuum. Actually, the group velocity in materials can be zero, infinite or even negative; this happens when the light pulse is very narrow, i.e. when it includes a wide range of frequencies, or again when the frequency is near an absorption transition. In many (but not all) cases the group is found to widen substantially or even to split, making it difficult to define precisely the group maximum and thus its velocity. Many experiments have confirmed these predictions. For example, the group velocity in certain materials has been measured to be *ten times* that of light. The refractive index then is smaller than 1. However, in all these cases the group velocity is *not* the same as the signal speed.*

Challenge 990 ny

Ref. 508

What then is the best velocity describing signal propagation? The German physicist Arnold Sommerfeld** almost solved the main problem in the beginning of the twentieth century. He defined the signal velocity as the velocity v_{So} of the front slope of the pulse,

* In quantum mechanics, Schrödinger proved that the velocity of an electron is given by the group velocity of its wavefunction. Therefore the same discussion reappeared in quantum theory, as we will find out in the second part of the mountain ascent.

** Arnold Sommerfeld (b. 1868 Königsberg, d. 1951 München) was a central figure in the spread of special

Ref. 507 as shown in Figure 252. The definition cannot be summarized in a formula, but it does have the property that it describes signal propagation for practically all experiments, in particular those in which the group and phase velocity are larger than the speed of light. When studying its properties, it is found that for no material Sommerfeld's signal velocity is larger than the speed of light in vacuum.

Sometimes it is conceptually easier to describe signal propagation with the help of the energy velocity. As mentioned before, every signal transports energy. The *energy velocity* v_{en} is defined as the ratio between the power flow density $\mathbf{P}$, i.e. the Poynting vector, and the energy density W, both taken in the direction of propagation. For electromagnetic fields – the only ones fast enough to be interesting for eventual superluminal signals – this ratio is

$$\mathbf{v}_{en} = \frac{\mathrm{Re}(\mathbf{P})}{W} = \frac{2c^2\,\mathbf{E}\times\mathbf{B}}{\mathbf{E}^2 + c^2\mathbf{B}^2} \,. \tag{442}$$

However, like in the case of the front velocity, also in the case of the energy velocity we have to specify if we mean the energy transported by the main pulse or by the front. In vacuum, neither is ever larger than the speed of light.* (In general, the energy velocity in Ref. 507 matter has a value slightly different from Sommerfeld's signal velocity.)

In recent years, the progress in light detector technology, allowing to detect even the tiniest energies, has forced people to take the fastest of all these energy velocities to describe signal velocity. Using detectors with the highest possible sensitivity we can use as signal the first point of the wave train whose amplitude is different from zero, i.e. the first tiny amount of energy arriving. This point's velocity, conceptually similar to Sommerfeld's signal velocity, is commonly called the *front velocity*, or, to distinguish it even Challenge 991 n more clearly from Sommerfeld's case, the *forerunner velocity*. It is simply given by

$$v_{fr} = \lim_{\omega\to\infty} \frac{\omega}{k} \,. \tag{443}$$

The forerunner velocity is *never* larger than the speed of light in vacuum, even in materials. In fact it is precisely c, because for extremely high frequencies, the ratio ω/k is independent of the material, and vacuum properties take over. The forerunner velocity is the true signal velocity or the *true velocity of light*. Using it, all discussions on light speed become clear and unambiguous.

Challenge 992 n To finish this section, here are two challenges. Which of all the velocities of light is measured in experiments determining the velocity of light, e.g. when light is sent to the Moon and reflected back? And now a more difficult one: why is the signal speed of light Challenge 993 n slower inside matter, as all experiments show?

---

and general relativity, of quantum theory, and of their applications. Professor in Munich, an excellent teacher and text book writer, he worked on atomic theory, on the theory of metals, on electrodynamics, and was the first to understand the importance and the mystery around 'Sommerfeld's famous fine structure constant.'

* Signals not only carry energy, they also carry negative entropy ('information'). The entropy of a transmitter increases during transmission. The receiver decreases in entropy (but less than the increase at the transmitter, Ref. 509 of course).

Note that the negative group velocity implies energy transport against the propagation velocity of light. Ref. 510 This is possible only in *energy loaded* materials.

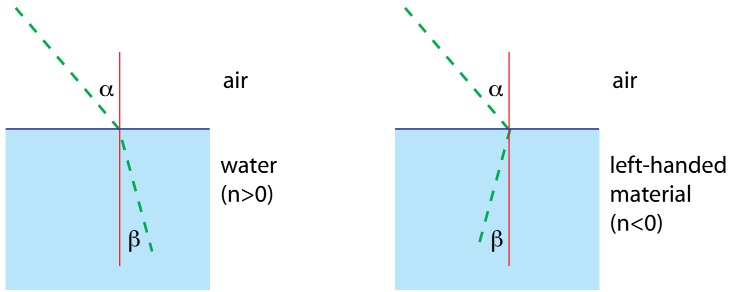


Figure 253 Positive and negative index of refraction

200 years too late: negative refraction indices

In 1968 the Soviet physicist Victor Veselago made a strange prediction: the index of refraction could have *negative* values without invalidating any known law of physics. A negative index of means that a beam is refracted on the same side, as shown in Figure 253.

In 1996, John Pendry and his group proposed ways to realize such materials. In 2000, a first experimental confirmation for microwaves was published, but met with strong Ref. 512 disbelief. In 2002 the debate was in full swing. It was argued that negative refraction indices imply speeds larger than that of light and are only possible for either phase velocity or group velocity, but not for the energy or true signal velocity. The conceptual problems would arise only because in some physical systems the refraction angle for phase motion and for energy motion differ.

Ref. 513 Today, the consensus is the following: a positive index of refraction smaller than one is impossible, as it implies an energy speed larger than one. A negative index of refraction, however, is possible, if it is smaller than −1. Negative values have indeed been frequently observed; the corresponding systems are being extensively explored all over the world. The materials showing this property are called *left-handed*. The reason is that the vectors of the electric field, the magnetic field and the wave vector form a left handed triplet, in contrast to vacuum and most usual materials, where the triplet is right-handed. Such materials consistently have negative magnetic permeability and negative dielectric coef- Ref. 511 ficient (permittivity).

Left-handed materials have negative phase velocities, i.e., phase velocity opposed to the energy velocity, they show reversed Doppler effects and yield obtuse angles in the Çerenkov effect (emitting Çerenkov radiation in the backward instead of the forward direction).

But most intriguing, negative refraction materials are predicted to allow the construc- Ref. 514 tion of lenses that are completely flat. In addition, in the year 2000, John Pendry got the attention of the whole physics community world-wide by predicting that lenses made with such materials, in particular for $n = -1$, would be *perfect*, thus beating the usual diffraction limit. This would happen because such a lens also images the evanescent parts of the waves, by amplifying them accordingly. First experiments seem to confirm the Ref. 511 prediction. The topic is still in full swing.

Challenge 994 ny Can you explain how negative refraction differs from diffraction?

Signals and predictions

When somebody reads a text through the phone to a neighbour who listens to it and maybe repeats it, we speak of communication. For any third person, the speed of communication is always smaller than the speed of light. But if the neighbour already knows the text, he can say it without waiting to hear the readers' voice. To the third observer such a situation looks like faster than light (superluminal) communication. Prediction can thus *mimic* communication, and in particular, it can mimic faster than light communication. Such a situation has been demonstrated most spectacularly in 1994 by Günter Nimtz, who seemingly transported music – all music is predictable for short time scales – through a 'faster-than-light' system. To distinguish between the two situations, we note that in the case of prediction, no energy transport takes place, in contrast to the case of communication. In other words, the definition of a signal as a transport of information is not as useful and clear-cut as the definition of a signal as *transport of energy*. In the mentioned experiment, no energy was transported faster than light. The same distinction between prediction on one hand and signal or energy propagation on the other hand will be used later on to clarify some famous experiments in quantum mechanics.

Ref. 515

> If the rate at which physics papers are being published continues to increase, physics journals will soon be filling library shelves faster than the speed of light. This does not violate relativity since no useful information is being transmitted.

How does the world look when riding on a light beam?

This was the question the teenager Albert Einstein tried to answer.* The situation would have strange consequences.

- You would have no mirror image, like a vampire.
- Light would not be oscillating, but a static field.
- Nothing would move, like in the tale of sleeping beauty.

But also at speeds *near* the velocity of light observations would be interesting. You would

- see a lot of light coming towards one and almost no light from the sides or from behind; the sky would be blue/white in front and red/black in the back;
- observe that everything around happens very very slowly;
- experience the smallest dust particle as deadly bullet.

Challenge 995 ny

Can you think of more strange consequences? It is rather reassuring that our planet moves rather slowly through its environment.

Does the aether exist?

Gamma rays, light and radio waves are moving electromagnetic waves. All exist in empty space. What is oscillating when the light comes along? Maxwell himself called the 'medium' in which this happens the *aether*. The properties of the aether measured in experiments are listed in Table 47.

* He took the question from a book on the sciences by Aaron Bernstein which he read at that time.

Table 47 Experimental properties of (flat) vacuum and of the 'aether'

Physical property	Experimental value
permeability	$\mu_0 = 1.3\,\mu\text{H/m}$
permittivity	$\varepsilon_0 = 8.9\,\text{pF/m}$
wave impedance/resistance	$Z_0 = 376.7\,\Omega$
conformal invariance	applies
spatial dimensionality	3
topology	R^3
mass and energy content	not detectable
friction on moving bodies	not detectable
motion relative to space-time	not detectable

Page 1061
Ref. 516
Challenge 996 n

Of course, the values of the permeability and the permittivity of vacuum are related to the definition of the units henry and farad. The last item of the table is the most important: despite intensive efforts, nobody has been able to detect any *motion* of the aether. In other words, even though the aether supposedly oscillates, it does not move. Together with the other data, all these results can be summarized in one sentence: there is no way to distinguish the aether from the vacuum: both are one and the same.

Sometimes it is heard that relativity or certain experiments show that the aether does not exist. That is not fully correct. In fact, experiments show something more important. All the data show that the aether is indistinguishable from the vacuum. Of course, if we use the change of curvature as definition for motion of the vacuum, vacuum *can* move, as we found out in the section on general relativity; but aether still remains indistinguishable from it.*

Ref. 517

Later we will even find out that the ability of the vacuum to allow the propagation of light and its ability to allow the propagation of particles are equivalent: both require the same properties. Therefore the aether remains *indistinguishable* from vacuum in the rest of our walk. In other words, the aether is a superfluous concept; we drop it from our walk from now on. Despite this result, we have not finished the study of the vacuum yet; it will keep us busy for a long time, starting with the intermezzo following this chapter. Moreover, quite a few of the aspects in Table 47 will require some amendments later on.

Challenges and curiosities about light

How to prove you're holy

Ref. 517

* In fact, the term 'aether' has been used as an expression for several different ideas, depending on the author. First of all it was used for the idea that vacuum is not empty, but *full*; secondly, that this fullness can be described by *mechanical models*, such as gears, little spheres, vortices, etc.; thirdly, it was imagined that vacuum is *similar to matter*, being made of the same substrate. Interestingly, some of these issues will reappear in the third part of our mountain ascent.

Figure 255 A limitation of the eye

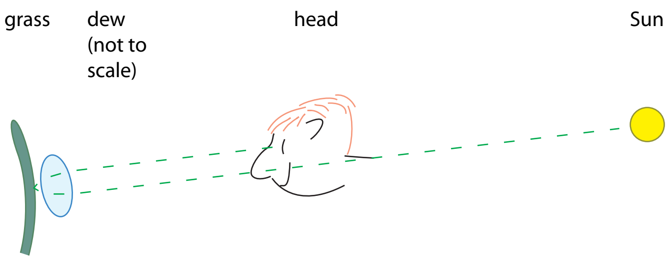


Figure 254 The path of light for dew on grass responsible for the aureole

Light reflection and refraction are responsible for many effects. The originally Indian symbol of holiness, now used throughout most of the world, is the *aureole*, also called *halo* or *Heiligenschein*, a ring of light surrounding the head. You can easily observe it around your own head. It is sufficient to get up early in the morning and to look into the wet grass while turning your back to the Sun. You will see an aureole around your shadow.

The effect is due to the morning dew on the grass, which reflects back the light mainly into the direction of the light source, as shown in the figure. The fun part is that if you do this in a group, you see the aureole only around *your own* head. Ref. 537

Retroreflective paint works in the same way; it contains tiny glass spheres which play the role of the dew. A large surface of retroreflective paint, a traffic sign for example, can also show your halo, if the light source is sufficiently far away. Also the so-called 'glow' of the eyes of cats at night is due to the same effect; it is visible only if you look at the cat with a light source in your back. By the way, does a Cat's-eye work like a cat's eye? Ref. 538 Challenge 997 n

Do we see what exists?

Sometimes we see *less* than there is. Close the left eye, look at the white spot in Figure 255, approach the page slowly to your eye, and pay attention to the middle lines. At a distance of about 15 to 20 cm the middle line will seem uninterrupted. Why? Challenge 998 n

On the other hand, sometimes we see *more* than there is, as Figures 256 and 257 show. They show so-called *Hermann lattices*, named after their discoverer.* These optical illu-

* Ludimar Herrmann (1838–1914), Swiss Physiologist. The lattices are often falsely called 'Hering lattices'

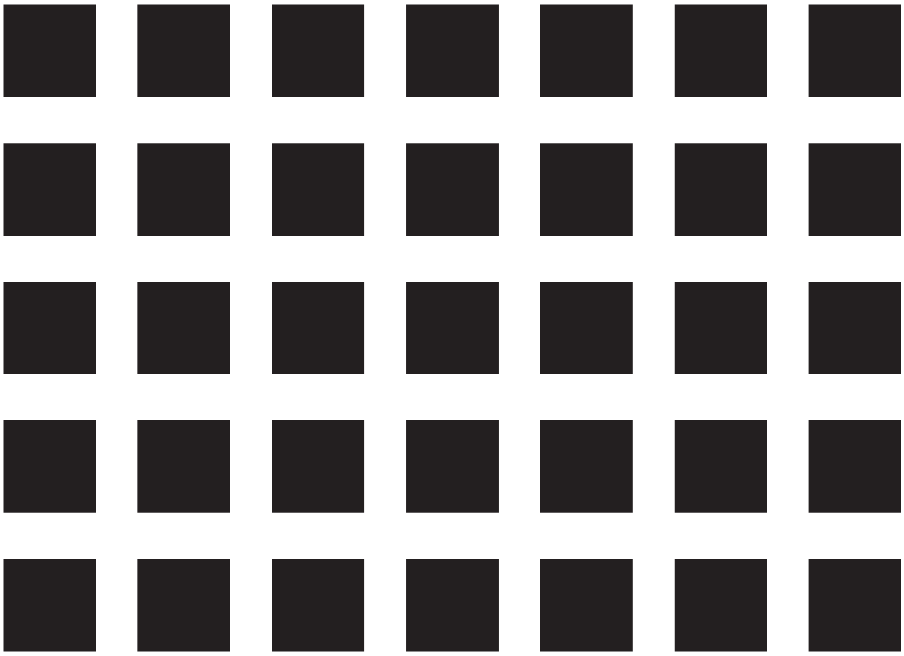


Figure 256 What is the shade of the crossings?

sions can even be used to determine how many light sensitive cells in the retina are united to one signal pathway towards the brain. The illusions are angle dependent because this number is also angle dependent. The lattice of Figure 257, discovered by Elke Lingelbach in 1995, is especially striking. Variations of these lattices are now used to understand the mechanisms at the basis of human vision. Ref. 539

Our eyes also see things *differently:* the retina sees an *inverted* image of the world. There is a simple method to show this, due to Helmholtz.* You only need a needle and a piece of paper, e.g. this page of text. Use the needle to make two holes inside the two letters 'oo'. Then keep the page as near to your eye as possible, look through the two holes towards the wall, keeping the needle vertical, a few centimetres behind the paper. You will see two images of the needle. If you now cover the *left* hole with your finger, the *right* needle will disappear, and vice versa. This shows that the image inside the eye, on the retina, is inverted. Are you able to complete the proof? Challenge 999 ny

Another reason that we do not see a complete image of nature is that the eye has a limited sensitivity. This sensitivity peaks around 560 nm; outside the red and the violet, the eye does not detect radiation. We thus see only part of nature. Infrared photographs

after the man who made Hermann's discovery famous.

* See Hermann von Helmholtz, *Handbuch der physiologischen Optik*, 1867. This famous classic is available in English as *Handbook of Physiological Optics*, Dover, 1962. The Prussian physician, physicist and science politician born as Hermann Helmholtz (b. 1821 Potsdam, d. 1894) was famous for his works on optics, on acoustics, electrodynamics, thermodynamics, epistemology and geometry. He founded several physics institutions across Germany. He was one of the first to propagate the idea of conservation of energy. His other important book, *Die Lehre von den Tonempfindungen*, published in 1863, describes the basis of acoustics, and like the handbook, is still worth to be read.

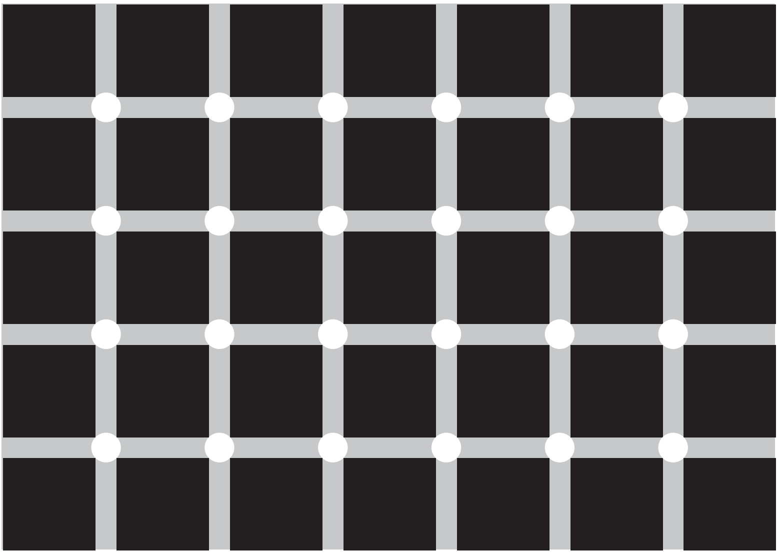


Figure 257 The Lingelbach lattice: do you see white, grey, or black dots?

of nature are often so interesting because they show us something which usually remains hidden. Every expert of motion should also know that the sensitivity of the eye does *not* correspond to the brightest part of sunlight. This myth is spread around the world by numerous textbooks copying from each other. Depending on whether frequency or wavelength or wavelength logarithm is used, the solar spectrum peaks at 500 nm, 880 nm or 720 nm. They eye's spectral sensitivity, like the completely different sensitivity of birds or frogs, is due to the chemicals used for detection; in short, the human eye can only be understood by a careful analysis of its particular evolution history. Ref. 540

In summary, we thus have to be careful when maintaining that seeing means observing. Examples such as these should make one ponder whether there could be other limitations of our senses which are less evident. And our walk will indeed uncover quite a few more.

Figure to be drawn

Figure 258 Eyes see inverted images

How does one make pictures of the inside of the eye?

The most beautiful pictures so far of a *living* human retina, such as that of Figure 259, were made by the group of David Williams and Austin Roorda at the University at Rochester in New York. They used adaptive optics, a technique which changes the shape of the imaging lens in order to compensate for the shape variations of the lens in the human eye.* Ref. 541

* Nature uses another trick to get maximum resolution: the eye continuously performs small movements, called *micronystagmus*. The eye continuously oscillates around the direction of vision with around 40 to

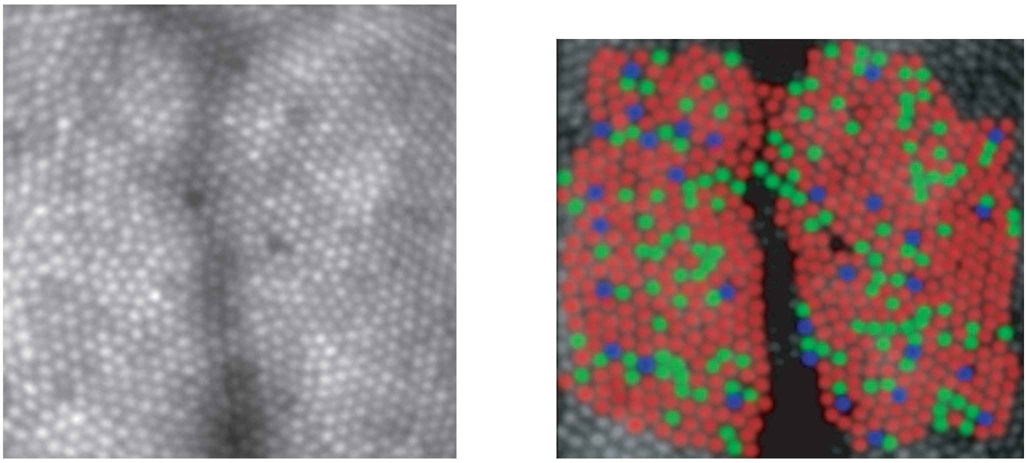


Figure 259 A high quality photograph of a live human retina, including a measured (false colour) indication of the sensitivity of each cone cell

The eyes see colour by averaging the intensity arriving at the red, blue and green sensitive cones. This explains the possibility, mentioned above, to get the same impression of colour, e.g. yellow, by a pure yellow laser beam, or by the mixture of red and green light. Page 531

But if the light is focussed onto one cone only, the eye makes mistakes. If, using this adaptive optics, a red laser beam is focussed such that it hits a green cone only, a strange thing happens: even though the light is *red*, the eye sees a *green* colour!

By the way, Figure 259 is quite puzzling. In the human eye, the blood vessels are located *in front* of the cones. Why don't they appear in the picture? And why don't they disturb us in everyday life? (The picture does not show the other type of sensitive light cells, the *rods*, because the person was in ambient light; rods come to the from of the retina only at dark, and then produce black and white pictures.

Challenge 1000 n

Amongst mammals, only primates can see *colours*. Bulls for example, don't; they cannot distinguish red from blue. On the other hand, the best colour seers overall are the birds. They have cone receptors for red, blue, green, UV and depending on the bird, for up to three more sets of colours. A number of birds (but not many) also have a better eye resolution than humans. Several birds also have a faster temporal resolution: humans see continuous motion when the images follow with 30 to 70 Hz (depending on the image content); some insects can distinguish images up to 300 Hz.

How does one make holograms and other 3-d images?

Our sense of sight gives us the impression of depth mainly due to three effects. First of all, the two eyes see different images. Secondly, the images formed in the eyes are position dependent. Thirdly, our eye needs to focus differently for different distances.

A simple photograph does not capture any of the three effects. A photograph corresponds to the picture taken by one eye, from one particular spot and at one particular focus. In fact, all photographic cameras are essentially copies of a single and static eye.

Challenge 1001 e

50 Hz. The motion is also used to allow the cells in the retina to recharge.

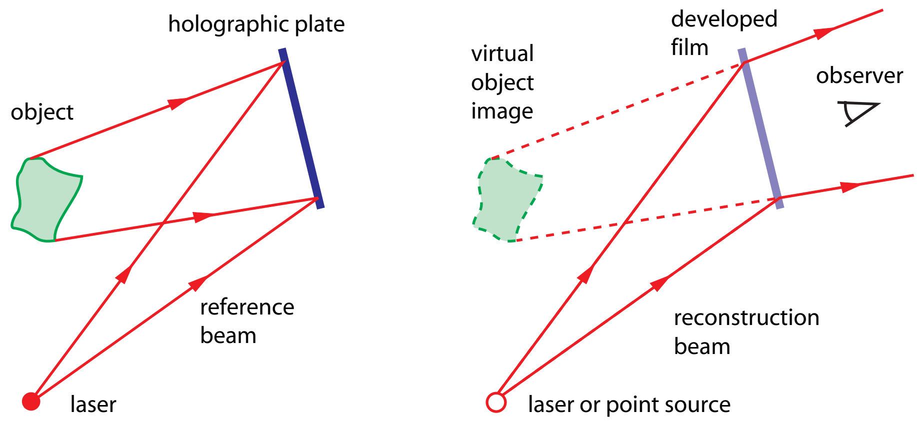


Figure 260 The recording and the observation of a hologram

Any system wanting to produce the perception of depth must include at least one of the three effects just mentioned. In all systems so far, the third and weakest effect, varying focus with distance, is never used, as it is too weak. Stereo photography and virtual reality systems extensively use the first effect by sending two different images to the eyes. Also certain post cards and computer screens are covered by thin cylindrical lenses which allow to send two different images to the two eyes, thus generating the same impression of depth.

But obviously the most spectacular effect is obtained whenever position dependent images can be created. Some virtual reality systems mimic this effect by attaching a sensor to the head, and creating computer–generated images which depend on this position. However, such systems are not able to reproduce actual situations and thus pale when compared to the impression produced by holograms.

Holograms reproduce all what is seen from any point of a region of space. A *hologram* is thus a stored set of position dependent pictures of an object. It is produced by storing amplitude *and phase* of the light emitted by an object. To achieve this, the object is illuminated by a *coherent* light source, such as a laser, and the interference pattern is stored. Illuminating the developed photographic film by a coherent light source then allows to see a full three-dimensional image. In particular, due to the reproduction of the situation, the image floats in free space. Holograms were developed in 1947 by the Hungarian physicist Dennis Gabor (1900–1979), who received the 1971 Nobel prize in physics for this work.

Holograms can be made to work in reflection or transmission. The simplest holograms use only one wavelength. Most coloured holograms are rainbow holograms, showing false colours that unrelated to the original objects. Real colour holograms, made with three different lasers, are rare but possible.

Is it possible to make *moving* holograms? Yes; however, the technical set-ups are still extremely expensive. So far, they only exist in a few laboratories and cost millions of euros. By the way, can you describe how you would distinguish a moving hologram from a real body, if you ever met one, without touching it? In fact, there is no way that holograms of people can walk around and frighten real people. Holograms look too much like the

Challenge 1002 n

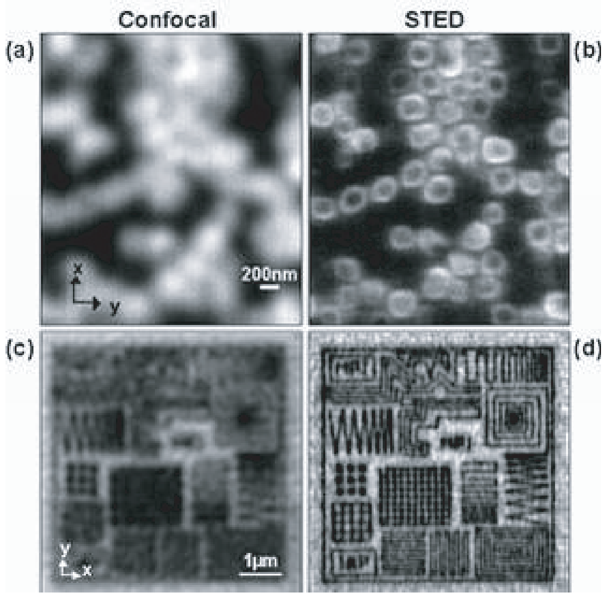


Figure 261 Sub-wavelength optical microscopy using stimulated emission depletion (© MPI für biophysikalische Chemie/Stefan Hell)

ghosts shown in many moving pictures.

Imaging

Producing images is an important part of modern society. The quality of taken and displayed images depends on the smart use of optics, electronics computers and materials science. Despite the long experience in this domain, there are still new results. For example, the techniques of producing images with resolution smaller than the wavelength of light has made fast progress in the recent years.

A recent technique, called *stimulated emission depletion microscopy*, allows spot sizes of molecular sizes. The conventional diffraction limit for microscopes is

$$d \geqslant \frac{\lambda}{2n\ \sin\alpha}\ , \tag{444}$$

where λ is the wavelength, n the index of refraction and α is the angle of observation. The new technique, a special type of fluorescence microscopy developed by Stefan Hell, modifies this expression to

$$d \geqslant \frac{\lambda}{2n\ \sin\alpha;\sqrt{I/I_{\text{sat}}}}\ , \tag{445}$$

so that a properly chosen saturation intensity allows to reduce the diffraction limit to arbitrary low values. So far, light microscopy with resolution of 16 nm has been performed. Ref. 518 This and similar techniques should become commonplace in the near future.

16. Charges are discrete – the limits of classical electrodynamics

Several remarks have already mentioned one of the most important results of physics: *electric charge is discrete*. Charge does not vary continuously, but changes in fixed steps. Not only does nature show a smallest value of entropy and smallest amounts of matter; nature also shows a smallest charge. Electric charge is quantized.

In metals, the quantization of charge is noticeable in the flow of electrons. In electrolytes, i.e. electrically conducting liquids, the quantization of charge appears in the flow of charged atoms, usually called *ions*. All batteries have electrolytes inside; also water is an electrolyte, though a poorly conducting one. In plasmas, like fire or fluorescent lamps, both ions and electrons move and show the discreteness of charge. Also in radiation – from the electron beams inside TVs, the channel rays formed in special glass tubes, the cosmic radiation up to radioactivity – charges are quantized.

In all known experiments, the same smallest value for charge change is found. The result is

$$\Delta q \geqslant e = 1.6 \times 10^{-19}\,\mathrm{C}\ . \tag{446}$$

In short, like all flows in nature, also the flow of electricity is due to a flow of discrete particles.

A smallest charge change has a simple implication: classical electrodynamics is *wrong*. A smallest charge implies that no infinitely small test charges exist. But such infinitely small test charges are necessary to define the electric and the magnetic field. The limit on charge size also implies that there is no correct way to define an instantaneous electric current, and as a consequence, that the values of electric and magnetic field are always somewhat fuzzy. Maxwell's evolution equations are thus only approximate.

We will study the main effects of the discreteness of charge in the part on quantum theory.702 Only few effects of the quantization of charge can be treated in classical physics. An instructive example is the following.

How fast do charges move?

Challenge 1003 n

In vacuum, such as inside a colour television, charged particles accelerated by a tension of 30 kV move with a third of the speed of light. In modern particle accelerators charges move so rapidly that their speed is indistinguishable from that of light for all practical purposes.

Inside metals, electric signals move with speeds of the order of the speed of light. The precise value depends on the capacity and impedance of the cable and is usually in the range $0.3c$ to $0.5c$. This high speed is due to the ability of metals to easily take in arriving charges and to let others depart. The ability for rapid reaction is due to the high mobility of the charges inside metals, which in turn is due to the small mass and size of these charges, the electrons.

The high signal speed in metals appears to contradict another determination. The drift speed of the electrons in a metal wire obviously obeys

$$v = \frac{I}{Ane} \tag{447}$$

where I is the current, A the cross section of the wire, e the charge of a single electron and n the number density of electrons. The electron density in copper is $8.5 \cdot 10^{28}\,\mathrm{m}^{-3}$. Using a typical current of 0.5 A and a typical cross section of a square millimetre, we get a drift speed of 0.37 μm/s. In other words, electrons move a thousand times slower than ketchup inside its bottle. Worse, if a room lamp would use direct current instead of alternate current, the electrons would take several days to get from the switch to the bulb! Nevertheless, the lamp goes on or off almost immediately after the switch is activated. Similarly, the electrons from an email transported with direct current would arrive much later than a paper letter sent at the same time; nevertheless, the email arrives quickly. Are you able to explain the apparent contradiction between drift velocity and signal velocity?

Challenge 1004 n

Inside liquids, charges move with a different speed than inside metals, and their charge to mass ratio is also different. We all know that from direct experience. Our *nerves* work by using electric signals and take (only) a few milliseconds to respond to stimuli, even though they are metres long. A similar speed is observed inside semiconductors and inside batteries. In all these systems, moving charge is transported by *ions*; they are charged atoms. Ions, like atoms, are large and composed entities, in contrast to the tiny electrons.

In other systems, charges move both as electrons and as ions. Examples are neon lamps, fire, plasmas or the Sun. Inside atoms, electrons behave even more strangely. One tends to think that they orbit the nucleus (as we will see later) at rather high speed, as the orbital radius is so small. However, it turns out that in most atoms many electrons do not orbit the nucleus at all. The strange story behind atoms and their structure will be told in the second part of our mountain ascent.

Page 704

Challenges and curiosities about charge discreteness

Challenge 1005 n

▪ How would you show experimentally that electrical charge comes in smallest chunks?

▪ The discreteness of charge implies that one can estimate the size of atoms by observing galvanic deposition. How?

Challenge 1006 ny

▪ Cosmic radiation consists of charged particles hitting the Earth. (We will discuss it in more detail later on.) Astrophysicists explain that these particles are accelerated by the magnetic fields around the galaxy. However, the expression of the Lorentz acceleration shows that magnetic fields can only change the direction of the velocity of charges, not its magnitude. How can nature get acceleration nevertheless?

Page 836 Ref. 519 Challenge 1007 ny

▪ What would be the potential of the Earth in volt if we could take all the electrons of a drop of water away?

Challenge 1008 n

▪ When an voltage is applied to a resistor, how long does it take until the end value of the current, given by Ohm's 'law', is reached? The first to answer this question was Paul Drude.* in the years around 1900. He reasoned that when the current is switched on, the speed v of an electron increases as $v = (eE/m)t$, where E is the electrical field, e the charge and m the mass of the electron. Drude's model assumes that the increase of electron speed stops when the electron hits an atom, loses its energy and starts again to

* Paul Karl Ludwig Drude (1863–1906), German physicist. A result of his electron gas model of metals was the prediction, roughly correct, that the ratio between the thermal conductivity and the electronic conductivity at a given temperature should be the same for all metals. Drude also introduced c as symbol for the speed of light.

be accelerated. Drude deduced that the average time τ up to the collision is related to the specific resistance by

Challenge 1009 ny

$$\rho = \frac{2m}{\tau e^2 n} \tag{448}$$

with n being the electron number density. Inserting numbers for copper ($n = 10.3 \cdot 10^{28}\,/\mathrm{m}^{-3}$ and $\rho = 0.16 \cdot 10^{-7}\,\Omega\mathrm{m}$), one gets a time $\tau = 42\,\mathrm{ps}$. This time is so short that the switch-on process can usually be neglected.

17. Electromagnetic effects and challenges

Classical electromagnetism and light are almost endless topics. Some aspects are too beautiful to be missed.

▪ Since light is a wave, something must happen if it is directed to a hole smaller than its wavelength. What happens exactly?

Challenge 1010 n

▪ Electrodynamics shows that light beams always push; they never pull. Can you confirm that 'tractor beams' are impossible in nature?

Challenge 1011 e

▪ It is well known that the glowing material in light bulbs is tungsten wire in an inert gas. This was the result of a series of experiments which started by the grandmother of all lamps, namely the cucumber. The older generation knows that a pickled cucumber, when attached to the 230 V of the mains, glows in bright green light. (Be careful; the experiment is dirty and somewhat dangerous)

▪ If you calculate the Poynting vector for a charged up magnet – or simpler, a point charge near a magnet – you get a surprising result: the electromagnetic energy flows in circles around the magnet. How is this possible? Where does this angular momentum come from?

Challenge 1012 n

Worse, any atom is an example of such a system – actually of two such systems. Why is this effect not taken into account in calculations in quantum theory?

Ref. 520

▪ Ohm's law, the observation that for almost all materials the current is proportional to the voltage, is due to a school teacher. Georg Simon Ohm explored the question in great depth; at those times, such measurements were difficult to perform.* This has changed now. Recently, even the electrical resistance of single atoms has been measured: in the case of xenon it turned out to be about $10^5\,\Omega$. It was also found that lead atoms are ten times more conductive than gold atoms. Can you imagine why?

Ref. 521

Challenge 1013 ny

▪ The charges on two capacitors in series are not generally equal, as naive theory states. For perfect, leak-free capacitors the voltage ratio is given by the inverse capacity ratio $V_1/V_2 = C_2/C_1$, due to the equality of the electric charges stored. However, in practice this is only correct for a few up to a few dozen minutes. Why?

Ref. 522

Challenge 1014 n

▪ Does it make sense to write Maxwell's equations in vacuum? Both electrical and magnetic fields require charges in order to be measured. But in vacuum there are no charges at all. In fact, only quantum theory solves this apparent contradiction. Are you able to imagine how?

Challenge 1015 d

* Georg Simon Ohm (b. 1789 Erlangen, d. 1854 München), bavarian school teacher and physicist. His efforts were recognized only late in his life, and he eventually was promoted to professor at the University in München. Later the unit of *electrical resistance*, the proportionality factor between voltage and current, was named after him.

Ref. 523

▪ Grass is usually greener on the other side of the fence. Can you give an explanation based on observations for this statement?

Challenge 1016 n

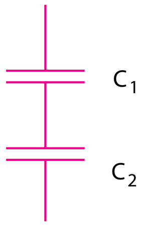


Figure 262 Capacitors in series

▪ 'Inside a conductor there is no electric field.' This statement is often found. In fact the truth is not that simple. First of all, a *static* field or a *static* charge on the metal surface of a body does not influence fields and charges inside it. A closed metal surface thus forms a shield against electric fields. Can you give an explanation? In fact, a tight metal layer is not required to get the effect; a cage is sufficient. One speaks of a *Faraday cage*.

Challenge 1017 n

The detailed mechanism allows you to answer the following question: do Faraday cages for gravity exist? Why?

For *moving* external fields or charges, the issue is more complex. Fields due to accelerated charges – radiation fields – decay exponentially through a shield. Fields due to charges moving at constant speed are strongly reduced, but do not disappear. The reduction depends on the thickness and the resistivity of the metal enclosure used. For sheet metal, the field suppression is very high; it is not necessarily high for metal sprayed plastic. Such a device will not necessarily survive a nearby lightning bolt.

Ref. 524

In practice, there is no danger if a lightning hits an aeroplane or a car, as long they are made of metal. (There is one movie on the internet of a car hit by a lightning; the driver does not even notice.) However, if your car is hit by lightning in dry weather, you should wait a few minutes before leaving it. Can you imagine why?

Faraday cages also work the other way round. (Slowly) changing electric fields changing inside a Faraday cage are not felt outside. For this reason, radios, mobile phones and computers are surrounded by boxes made of metal or metal-sprayed plastics. The metal keeps the so-called *electromagnetic smog* to a minimum.

There are thus three reasons to surround electric appliances by a grounded shield: to protect the appliance from outside fields, to protect people and other machines from electromagnetic smog, and to protect people against the mains voltage accidentally being fed onto the box (for example, when insulation fails). In high precision experiments, these three functions can be realized by three separate cages.

For purely magnetic fields, the situation is more complex. It is quite difficult to shield the inside of a machine from outside magnetic fields. How would you do it? In practice one often uses layers of so-called *mu-metal*; can you guess what this material does?

Challenge 1018 ny

▪ The *electric polarizability* is the property of matter responsible for the deviation of water flowing from a tap by a charged comb. It is defined as the strength of electric dipole induced by an applied electric field. The definition simply translates the observation that many objects acquire charges when an electric field is applied. Incidentally, how precisely combs get charged when rubbed, a phenomenon called *electrification*, is still one of the mysteries of modern science.

Page 479

▪ A pure magnetic field cannot be transformed into a pure electric field by change of observation frame. The best that can be achieved is a state similar to an equal mixture of magnetic and electric fields. Can you provide an argument elucidating this relation?

Challenge 1019 ny

Ref. 525

▪ Researchers are trying to detect tooth decay with the help of electric currents, using the observation that healthy teeth are bad conductors, in contrast to teeth with decay. How would you make use of this effect in this case? (By the way, it might be that the totally unrelated technique of imaging with terahertz waves could yield similar results.)

Challenge 1020 ny

▪ A team of camera men in the middle of the Sahara were using battery driven electrical equipment to make sound recordings. Whenever the microphone cable was a few tens of metres long, they also heard a 50 Hz power supply noise, even though the next power supply was thousands of kilometres away. An investigation revealed that the high voltage lines in Europe lose a considerable amount of power by irradiation; those 50 Hz waves are reflected by the ionosphere around the Earth and thus can disturb recording in the middle of the desert. Can you estimate whether this observation implies that living directly near a high voltage line is dangerous?

Challenge 1021 ny

▪ On certain high voltage cables leading across the land scape, small neon lamps shine when the current flows. How is that possible?

Challenge 1022 ny

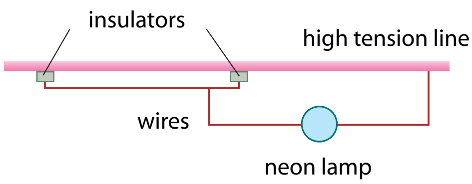


Figure 263 Small neon lamps on a high voltage cable

▪ When two laser beams cross at a small angle, one can form light pulses which seem to move faster than light. Does this contradict special relativity?

Ref. 528
Challenge 1023 n

▪ It is said that astronomers have telescopes so powerful that they can see whether somebody would light a match on a Moon. Can this be possible?

Challenge 1024 ny

▪ When solar plasma storms are seen on the Sun, astronomers first of all phone the electricity company. They know that about 24 to 48 hours later, the charged particles ejected by the storms will arrive on Earth, making the magnetic field on the surface fluctuate. Since power grids often have closed loops of several thousands of kilometres, additional electric currents are induced, which can make transformers in the grid overheat and then switch off. Then other transformers have to take over the additional power, which can lead to their overheating etc. Several times in the past, millions of people were left without electrical power due to solar storms. Today, the electricity companies avoid the problems by disconnecting the various grid sections, by avoiding large loops, by reducing the supply voltage to avoid saturation of the transformers and by disallowing load transfer from failed circuits to others.

▪ Is it really possible to see stars from the bottom of a deep pit or of a well even during daytime, as is often stated also in print?

Challenge 1025 n

▪ If the electric field is described as a sum of components of different frequencies, its so-called *Fourier components*, the amplitudes are given by

Ref. 526

$$\hat{\mathbf{E}}(k,t) = \frac{1}{(2\pi)^3/2} \int \mathbf{E}(x,t) e^{-i\mathbf{kx}} \, \mathrm{d}^3 x \qquad (449)$$

and similarly for the magnetic field. It then turns out that a Lorentz invariant quantity N, describing the energy per circular frequency ω, can be defined:

$$N = \frac{1}{8\pi} \int \frac{|\mathbf{E}(k,t)|^2 + |\mathbf{B}(k,t)|^2}{c|\mathbf{k}|} \, \mathrm{d}^3 k \qquad (450)$$

Challenge 1026 n

Can you guess what N is physically? (Hint: think about quantum theory.)

▪ Faraday discovered how to change magnetism into electricity, knowing that electricity could be transformed into magnetism. (The issue is subtle. Faraday's law is not the dual of Ampère's, as that would imply the use of magnetic monopoles; neither is it the re-

Figure 264 How natural colours (top) change for three types of colour blind: deutan, protan and tritan (© Michael Douma)

ciprocal, as that would imply the displacement current. But he was looking for a link and he found a way to relate the two observations – in a novel way, as it turned out.) Faraday also discovered how to transform electricity into light and into chemistry. He then tried to change gravitation into electricity. But he was not successful. Why not? Challenge 1027 ny

▪ Take an envelope, wet it and close it. After letting it dry for a day or more, open it in the dark. At the place where the two papers are being separated from each other, the envelope glows with a blue colour. Why? Challenge 1028 ny

▪ At high altitudes above the Earth, gases are completely ionized; no atom is neutral. One speaks of the *ionosphere*, as space is full of positive ions and free electrons. Even though both charges appear in exactly the same number, a satellite moving through the ionosphere acquires a negative charge. Why? How does the charging stop? Challenge 1029 n

▪ A capacitor of capacity C is charged with a voltage U. The stored electrostatic energy is $E = CU^2/2$. The capacitor is then detached from the power supply and branched onto an empty capacitor of the same capacity. After a while, the voltage obviously drops to $U/2$. However, the stored energy now is $C(U/2)^2$, which is half the original value. Where did the energy go? Challenge 1030 n

▪ Colour blindness was discovered by the great English scientist John Dalton (1766–1844) – on himself. Can you imagine how he found out? Challenge 1031 ny It affects, in all its forms, one in 20 men. In many languages, a man who is colour blind is called *daltonic*. Women are almost never daltonic, as the property is linked to defects on the X chromosome. Ref. 529 If you are colour blind, you can check which type you belong to with the help of Figure 264.

▪ Perfectly spherical electromagnetic waves are impossible in nature. Can you show this using Maxwell's equation of electromagnetism, or even without them? Challenge 1032 n

▪ Light beams, such as those emitted from lasers, are usually thought as lines. However, light beams can also be *tubes*. Tubular laser beams, or Bessel beams of high order, are used in modern research to guide plasma channels.

Figure 265 Cumulonimbus clouds from ground and from space (courtesy NASA)

Is lightning a discharge? – Electricity in the atmosphere

Looking carefully, the atmosphere is full of electrical effects. The most impressive electrical phenomenon we observe, the lightning, is now reasonably well understood. Inside a thunderstorm cloud, especially inside tall *cumulonimbus* clouds,* charges are separated by collision between the falling large 'graupel' ice crystals falling due to their weight and Ref. 531 the small 'hail' ice crystallites rising due to thermal upwinds. Since the collision takes part in an electric field, charges are separated in a way similar to the mechanism in the Page 481 Kelvin generator. Discharge takes place when the electric field becomes too high, taking a strange path influenced by ions created in the air by cosmic rays. It seems that cosmic rays are at least partly responsible for the zigzag shape of lightning.** Lighting have strange properties. First of all, they appear already at fields around 200 kV/m (at low altitude) instead of the 2 MV/m of normal sparks. Second, lightning emit radio pulses. Third, they Ref. 533 emit gamma rays. Russian researchers, from 1992 onwards explained all three effects with a newly discovered discharge mechanism. At length scales of 50 m and more, cosmic rays can trigger the appearance of lightning; the relativistic energy of these rays allows for a discharge mechanism that does not exist for low energy electrons. At relativistic energy, so-called runaway breakdown leads to discharges at much lower fields than usual laboratory sparks. The multiplication of these relativistic electrons also leads to the observed radio and gamma ray emission.

* Clouds have latin names. They were introduced in 1802 by the English explorer Luke Howard (1772–1864), who found that all clouds could be seen as variations of three types, which he called *cirrus*, *cumulus* Ref. 530 and *stratus*. He called the combination of all three, the rain cloud, *nimbus* (from latin 'big cloud'). Today's internationally agreed system has been slightly adjusted and distinguishes clouds by the height of their lower edge. The clouds starting above a height of 6 km are the cirrus, the cirrocumulus and the cirrostratus; those starting at a height between 2 and 4 km are the altocumulus, the altostratus and the nimbostratus; clouds starting below a height of 2 km are the stratocumulus, the stratus and the cumulus. The rain or thunder cloud, which crosses all heights, is today called cumulonimbus.

Ref. 532 ** There is no ball lightning even though there is a Physics Report about them. Ball lightning is one of the favourite myths of modern pseudo-science. Actually, they would exist if we lived in a giant microwave oven. To show this, just stick a toothpick into a candle, light the toothpick, and put it into (somebody else's) microwave at maximum power.

By the way, you have a 75 % survival chance after being hit by lightning, especially if your are fully wet, as in that case the current flows outside the skin. Usually, wet people who are hit lose all their clothes, as the evaporating water tears them off. Rapid resuscitation is essential to help somebody to recover after a hit.*

As a note, you might know how to measure the distance of a lightning by counting the seconds between the lightning and the thunder and multiplying by the speed of sound, 330 m/s; it is less well known that one can estimate the *length* of the lightning bolt by measuring the *duration* of the thunder, and multiplying by it the same factor.

In the nineteen nineties, more electrical details about thunderstorms became known. Airline pilots and passengers sometime see weak and coloured light emissions spreading from the top of thunderclouds. There are various types of such emissions, blue *jets* and mostly red *sprites* and *elves*, which are somehow due to electric fields between the cloud top and the ionosphere. The details are still under investigation, and the mechanisms are not yet clear.*

All these details are part of the electrical circuit around the Earth. This fascinating part of geophysics would lead us too far from the aim of our mountain ascent. But every physicist should know that there is a vertical electric field of between 100 and 300 V/m on a clear day, as discovered already in 1752. (Can you guess why it is not noticeable in everyday life? And why despite its value it cannot be used to extract large amounts of energy?) The field is directed from the ionosphere downwards to the ground; in fact the Earth is permanently charged negatively, and on clear weather current flows downwards through the clear atmosphere, trying to *discharge* our planet. The current of about 1 kA is spread over the whole planet; it is possible due to the ions formed by cosmic radiation. (The resistance between the ground and the ionosphere is about 200 Ω, so that the total voltage drop is about 200 kV.) At the same time, the Earth is constantly being *charged* by several effects, of which the most important one turns out to be the lightning. In other words, contrary to what one may think, lightning do not discharge the ground, they actually charge it up!** Of course, lightning does discharge the cloud to ground potential difference, but by doing so, it actually sends negative charge down to the Earth.

Challenge 1033 n

Using a few electrical measurement stations that measure the variations of the electrical field of the Earth it is possible to locate the position of all the lightning that come down towards the Earth in a given moment. Present research also aims at measuring the activity of the related electrical sprites and elves in this way.

Ref. 534

The ions in air play a role in the charging of thunderclouds via the charging of ice crystals and rain drops. In general, all small particles in the air are electrically charged. When aeroplanes and helicopters fly, they usually hit more particles of one charge than of the other. As a result, aeroplanes and helicopters are charged up during flight. When a

* If you are hit by lightning and survive, go to the hospital! Many people died three days later for failing to do so. A lightning stroke often leads to coagulation effects in the blood. These substances block the kidneys, and people die three days later because of kidney failure. The way to help is to have a dialysis treatment.

* For images, have a look at the interesting http://sprite.gi.alaska.edu/html/sprites.htm, http://www.fma-research.com/spriteres.htm and http://paesko.ee.psu.edu/Nature websites.

Challenge 1034 ny

** The Earth is thus charged to about −1 MC. Can you confirm this? To learn more about atmospheric currents, you may want to have a look at the popularizing review of US work by Edgar Bering, Arthur Few & James Benbrook, The global electric circuit, *Physics Today* **51**, pp. 24–30, October 1998, or the more technical overview by Edgar Bering, *Reviews of Geophysics* (supplement) 33, p. 845, 1995.

helicopter wants to save people on a raft in high sea, the rope pulling the people upwards must first be earthed by hanging it into the water; if this is not done, the people can die from electrical shock when they touch the rope, as happened a few times in the past.

The charges in the atmosphere have many other effects. Recent experiments have confirmed what has been predicted already in the early twentieth century: lightning emit X-rays. The confirmation is not easy though; it is necessary to put a detector near the lightning stroke. To achieve this, lightning has to be directed into a given region. This is possible using missiles pulling metal wires with them, of which the other end is attached to ground. These experimental results are now being summarized into a new description of lightning which also explains the red-blue sprites above thunderclouds. In particular, the processes also imply that inside clouds, electrons can be accelerated up to energies of a few MeV.

Ref. 535

Ref. 536

Why are sparks and lightning blue? This turns out to be a material property; the colour is given by the material that happens to be excited by the energy of the discharge, usually air. This excitation is due to the temperature of 30 kK inside the channel of a typical lightning stroke. For everyday sparks, the temperature is much smaller. Depending on the situation, the colour may arise from the gas between the two electrodes, such as oxygen or nitrogen, or it may due to the material evaporated from the electrodes by the discharge. For an explanation of such colours, like for the explanation of all material related colours, we need to wait for the next part of our walk.

But not only electric fields are dangerous. Also the time-varying electromagnetic fields can be. In 1997, with beautiful calm weather, a Dutch hot air balloon approached the powerful radio transmitter in Hilversum. After a few minutes near the antenna, the gondola suddenly detached from the balloon, killing all passengers inside.

An investigation team reconstructed the facts a few weeks later. In modern gas balloons the gondola is suspended by high quality nylon ropes. To avoid damage by lightning and in order to avoid electrostatic charging problems all these nylon ropes contain thin metal wires which form a large equipotential surface around the whole balloon. Unfortunately, in front of the radio transmitter these thin metal wires absorbed the radio energy from the transmitter, became red hot and melted the nylon wires. It was the first time that this was ever observed.

Does gravity make charges radiate?

We learned in the section on general relativity that gravitation has the same effects as acceleration. This means that a charge kept fixed at a certain height is equivalent to a charge accelerated by $9.8\,\mathrm{m/s^2}$, which would imply that it radiates electromagnetically, since all accelerated charges radiate. However, the world around us is full of charges at fixed heights, and there is no such radiation. How is this possible?

Ref. 542

The question has been a pet topic for many years. It turns out that the answer depends on whether the observer detecting the radiation is also in free fall or not, and on the precise instant this started to be the case.

In practice, gravity does not make electrical charges radiate.

Research questions

The classical description of electrodynamics is coherent and complete; nevertheless there are still many subjects of research. Here are a few of them.

The origin of magnetic field of the Earth, the other planets, the Sun and even of the galaxy is a fascinating topic. The way that the convection of fluids inside the planets generates magnetic fields, an intrinsically three dimensional problem, the influence of turbulence, of nonlinearities, of chaos etc. makes it a surprisingly complex question.

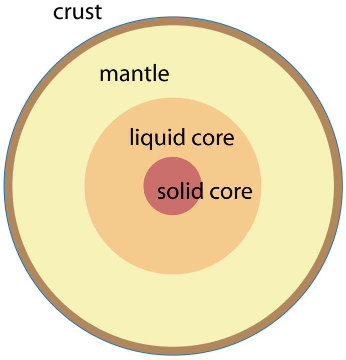


Figure 266 The structure of our planet

The details of the generation of the magnetic field of the Earth, usually called the *geodynamo*, began to appear only in the second half of the twentieth century, when the knowledge of the Earth's interior reached a sufficient level. The Earth's interior starts below the Earth's crust. Ref. 544 The *crust* is typically 30 to 40 km thick (under the continents), though thicker under high mountains and thinner near volcanoes or under the oceans. As already mentioned, the crust consists of large segments, the *plates*, which move with respect to each other. The Earth's interior is divided into the *mantle* – the first 2900 km from the surface – and the *core*. The core is made of a liquid *outer* core, 2300 km thick, and a solid *inner* core of 1215 km radius. (The temperature of the core is not well known; it is believed to be 6 to 7 kK. Challenge 1035 ny Can you find a way to determine it? The temperature might have decreased a few hundred kelvin during the last 3000 million years.)

The Earth's core consists mainly of iron which has been collected from the asteroids that have collided with the Earth during its youth. It seems that the liquid and electrically conducting outer core acts as a dynamo which keeps the magnetic field going. The magnetic energy comes from the kinetic energy of the outer core, which rotates with respect to the Earth's surface; the fluid can act as a dynamo because, apart from rotating, it also *convects* from deep inside the Earth to more shallow depths, driven by the temperature gradients between the hot inner core and the cooler mantle. Huge electric currents flow in complex ways through these liquid layers, maintained by friction, and create the magnetic field. Understanding why this field switches orientation at irregular intervals of between a few tens of thousands and a few million years, is one of the central questions. The answers are difficult; experiments are not yet possible, 150 years of measurements is a short time when compared to the last transition – about 730 000 years ago – and computer simulations are extremely involved. Since the field measurements started, the dipole moment of the magnetic field has steadily diminished, presently by 5% a year, and the quadrupole moment has steadily increased. Maybe we are heading towards a surprise. (By the way, the study of *galactic* magnetic fields is even more complex, and still at its beginning.)

Another important puzzle on electricity results from the equivalence of mass and energy. It is known from experiments that the size d of electrons is surely smaller than Ref. 545 10^{-22} m. This means that the electric field surrounding it has an energy content E given Challenge 1036 ny by at least

$$E_{\text{nergy}} = \frac{1}{2}\varepsilon_0 \int E^2_{\text{lectric field}}\, dV = \frac{1}{2}\varepsilon_0 \int_d^\infty \Big(\frac{1}{4\pi\varepsilon_o}\frac{q}{r^2}\Big)^2 4\pi r^2 dr$$
$$= \frac{q^2}{8\pi\varepsilon_o}\frac{1}{d} > 1.2\,\mu\text{J} . \qquad (451)$$

On the other hand, the *mass* of an electron, usually given as 511 keV/c^2, corresponds to an energy of only 82 fJ, ten million times *less* than the value just calculated. In other words, classical electrodynamics has considerable difficulty describing electrons. In fact, a consistent description of charged point particles within classical electrodynamics is not possible. This pretty topic receives only a rare – but then often passionate – interest nowadays, because the puzzle is solved in a different way in the upcoming, second part of our mountain ascent.

Ref. 546

Even though the golden days of materials science are over, the various electromagnetic properties of matter and their applications in devices do not seem to be completely explored yet. About once a year a new effect is discovered which merits to be included in the list of electromagnetic matter properties of Table 48. Among others, some newer semiconductor technologies will still have an impact on electronics, such as the recent introduction of low cost light detecting integrated circuits built in CMOS (complementary metal oxide silicon) technology.

Page 558

Building light sources of high quality has been a challenge for many centuries and still remains one for the future. Light sources which are intense, tunable, with large coherence length or sources which emit extreme wavelengths are central to many research pursuits. As an example among many, the first X-ray lasers have been built recently; however, they are several hundred metres in size and use modified particle accelerators. The construction of compact X-ray lasers is still many years away – if it is possible at all.

Electrodynamics and general relativity interact in many ways. Only a few cases have been studied up to now. They are important for black holes and for empty space. For example, it seems that magnetic fields increase the stiffness of empty space. Many such topics will appear in the future.

Ref. 543

But maybe the biggest challenge imaginable in classical electrodynamics is to decode the currents inside the brain. Is it possible to read our thoughts with an apparatus placed outside the head? One could start with a simple challenge: is it possible to distinguish the thought 'yes' from the though 'no' by measuring electrical or magnetic fields around the head? In other words, is mind-reading possible? Maybe the twenty-first century will give us a positive answer. If so, the team performing the feat will be instantly famous.

Challenge 1037 r

Levitation

We have seen that it is possible to move certain objects without touching them, using a magnetic or an electric field, or of course, using gravity. Is it also possible, without touching an object, to keep it fixed, floating in mid air? Does this type of rest exist?

It turns out that there are several methods to levitate objects. They are commonly divided into two groups: those which consume energy and those who do not. Among the methods consuming energy is the floating of objects on a jet of air or of water, the floating of objects through sound waves, e.g. on top of a siren, or through a laser beam coming from below, and the floating of conducting material, even of liquids, in strong radio-

Ref. 547

frequency fields. Levitation of liquids or solids by strong ultrasound waves is presently becoming popular in laboratories. All these methods give *stationary* levitation. Another group of energy consuming methods sense the way a body is falling and kick it up again in the right way via a feedback loop; these methods are *non-stationary* and usually use magnetic fields to keep the objects from falling. The magnetic train being built in Shanghai by a German consortium is levitated this way. The whole train, including the passengers, is levitated and then moved forward with electromagnets. It is thus possible, using magnets, to levitate many tens of tons of material.

Ref. 548
Ref. 549

For levitation methods which do *not* consume energy – all such methods are necessarily stationary – a well-known limitation can be found studying Coulomb's 'law' of electrostatics: no static, i.e. time-independent arrangement of electric fields can levitate a *charged* object in free space or in air. The same result is valid for gravitational fields and *massive* objects;* in other words, we cannot produce a local minimum of potential energy in the middle of a box using electric or gravitational fields. This impossibility is called *Earnshaw's theorem*. Speaking mathematically, the solutions of the Laplace equation $\Delta\varphi = 0$, the so-called *harmonic functions*, have minima or maxima only at the border, and never inside the domain of definition. (You proved this yourself on page 114.) The theorem can also be proved by noting that given a potential minimum in free space, Gauss' theorem for a sphere around that minimum requires that a source of the field be present inside, which is in contradiction with the original assumption.

Ref. 550

We can deduce that it is also impossible to use electric fields to levitate an electrically *neutral* body in air: the potential energy U of such a body, with volume V and dielectric constant ε, in an environment of dielectric constant ε_0, is given by

$$\frac{U}{V} = -\frac{1}{2}(\varepsilon - \varepsilon_0)E^2 \quad . \tag{452}$$

Challenge 1038 ny

Since the electric field E never has a maximum in the absence of space charge, and since for all materials $\varepsilon > \varepsilon_0$, there cannot be a minimum of potential energy in free space for a neutral body.**

In summary, using static electric or static gravitational fields it is impossible to keep an object from falling; neither quantum mechanics, which incorporates phenomena such as antimatter, nor general relativity, including phenomena such as black holes, change this basic result.

For static *magnetic* fields, the argument is analogous to electrical fields: the potential energy U of a magnetizable body of volume V and permeability μ in a medium with permeability μ_0 containing no current is given by

Challenge 1040 ny

$$\frac{U}{V} = -\frac{1}{2}\left(\frac{1}{\mu} - \frac{1}{\mu_0}\right)B^2 \tag{453}$$

* To the disappointment of many science-fiction addicts, this would also be true in case that negative mass would exist, as happens for charge. See also page 76. And even though gravity is not really due to a field, the result still holds in general.

** It is possible, however, to 'levitate' gas bubbles in liquids – 'trap' them to prevent them from rising would be a better expression – because in such a case the dielectric constant of the environment is higher than that of the gas. Can you find a liquid–gas combination where bubbles fall instead of rising; can you find one?

Ref. 551
Challenge 1039 ny

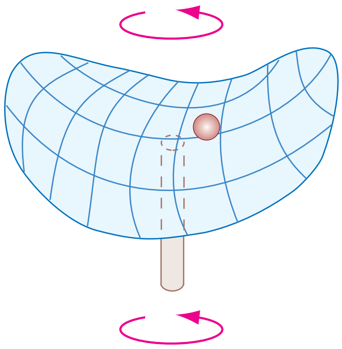


Figure 267 Trapping a metal sphere using a variable speed drill and a plastic saddle

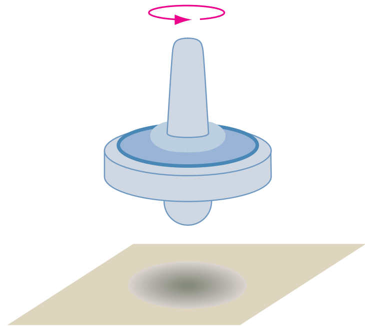


Figure 268 Floating 'magic' nowadays available in toy shops

and due to the inequality $\Delta B^2 \geqslant 0$, isolated maxima of a static magnetic field are not possible, only isolated minima. Therefore, it is impossible to levitate paramagnetic ($\mu > \mu_o$) or ferromagnetic ($\mu \gg \mu_0$) materials such as steel, including bar magnets, which are all attracted, and not repelled to magnetic field maxima.

Challenge 1041 ny

There are thus two ways to get magnetic levitation: levitating a diamagnet or using a time dependent field. Diamagnetic materials ($\mu < \mu_o$) can be levitated by static magnetic fields because they are attracted to magnetic field minima; the best-known example is the levitation of superconductors, which are, at least those of type I, perfects diamagnets ($\mu = 0$). Strong forces can be generated, and this method is also being tested for the levitation of passenger trains in Japan. In some cases, superconductors can even be *suspended* in mid-air, below a magnet. Single atoms with a magnetic moment are also diamagnets; they are routinely levitated this way and have also been photographed in this state.

Ref. 552 Ref. 549 Ref. 553

Also single neutrons, which have a magnetic dipole moment, have been kept in magnetic bottles in this way, until they decay. Recently, people have levitated pieces of wood, of plastic, strawberries, water droplets, liquid helium droplets as large as 2 cm, grasshoppers, fish and frogs (all alive and without any harm) in this way. They are, like humans, all made of diamagnetic material. Humans themselves have not yet been levitated, but the feat is being planned and worked on.

Ref. 554

Diamagnets levitate if $\nabla B^2 > 2\mu_0 \rho g/\chi$, where ρ is the mass density of the object and $\chi = 1 - \mu/\mu_0$ its magnetic susceptibility. Since χ is typically about 10^{-5} and ρ of order $1000\,\mathrm{kg/m^3}$, field gradients of about $1000\,\mathrm{T^2/m}$ are needed. In other words, levitation requires fields changes of 10 T over 10 cm, nowadays common for high field laboratory magnets.

Challenge 1042 ny

Finally, *time dependent* electrical or magnetic fields, e.g. periodic fields, can lead to levitation in many different ways without any consumption of energy. This is one of the methods used in the magnetic bearings of turbomolecular vacuum pumps. Also single charged particles, such as ions and electrons, are now regularly levitated with Paul traps and Penning traps. The mechanical analogy is shown in Figure 267.

Ref. 547 Ref. 547

Figure 268 shows a toy allowing to let one personally levitate a spinning top in mid air above a ring magnet, a quite impressive demonstration of levitation for anybody looking

Ref. 555

Ref. 556 at it. It is not hard building such a device oneself.

Even free electrons can be levitated, letting them float above the surface of fluid helium. Ref. 557 In the most recent twist of the science of levitation, in 1995 Stephen Haley predicted that the suspension height of small magnetic particles above a superconducting ring should be quantized. However, the prediction has not been checked by experiment yet.

For the sake of completeness we mention that the nuclear forces cannot be used for levitation in everyday life, as their range is limited to a few femtometres. However, we Page 858 will see later that the surface matter of the Sun is prevented from falling into the centre by these interactions; we could thus say that it is indeed levitated by nuclear interactions.

Matter, levitation and electromagnetic effects

Levitation used by magicians mostly falls into another class. When David Copperfield, a magician performing for the MTV generation at the end of the twentieth century, 'flies' during his performances, he does so by being suspended on thin fishing lines kept invis- Ref. 558 ible by clever lighting arrangements. In fact, if we want to be precise, we should count fishing lines, plastic bags, as well as every table and chair as levitation devices. (Journalists would even call them 'anti-gravity' devices.) Contrary to first impression, a hanging or lying object is not really in contact with the suspension, if we look at the critical points with a microscope.* More about this in the second part of our walk.

But if this is the case, why don't we fall through a table or through the floor? We started the study of mechanics by stating as key property of matter its *solidity*, i.e. the impossibility to have more than one body at the same place at the same time. But what is the origin of solidity? Again, we will be able to answer the question only in the part on quantum Page 783 mechanics, but we can collect the first clues already at this point.

Solidity is due to electricity. Many experiments show that matter is constituted of charged particles; indeed, matter can be moved and influenced by electromagnetic fields Ref. 559 in many ways. Over the years, material scientists have produced a long list of such effects, all of which are based on the existence of charged constituents. Can you find or imagine Challenge 1044 r a new one? For example, can electric charge change the colour of objects?

Table 48 Selected matter properties related to electromagnetism, showing among others the role it plays in the constitution of matter; at the same time a short overview of atomic, solid state, fluid and business physics

Property	Example	Definition
thermal radiation or heat radiation or incandescence	every object	temperature dependent radiation emitted by any macroscopic amount of matter
Interactions with charges and currents		
electrification	separating metals from insulators	spontaneous charging
triboelectricity	glass rubbed on cat fur	charging through rubbing
barometer light	mercury slipping along glass	gas discharge due to triboelectricity Ref. 560

Challenge 1043 ny * The issue is far from simple: which one of the levitation methods described above is used by tables or chairs?

Property	Example	Definition
insulation	air	no current flow below critical voltage drop
semiconductivity	diamond, silicon or gallium arsenide	current flows only when material is impure ('doped')
conductivity	copper, metals	current flows easily
superconductivity	niobium	current flows indefinitely
ionization	fire flames	current flows easily
localization (weak, Anderson)	disordered solids	resistance of disordered solids
resistivity, Joule effect	graphite	heating due to current flow
thermoelectric effects: Peltier effect, Seebeck effect, Thomson effect	ZnSb, PbTe, PbSe, BiSeTe, Bi_2Te_3, etc.	cooling due to current flow, current flow due to temperature difference, or due to temperature gradients
acoustoelectric effect	CdS	sound generation by currents, and vice versa
magnetoresistance	iron, metal multilayers	resistance changes with applied magnetic field Ref. 561
recombination	fire alarms	charge carriers combine to neutral atoms or molecules
annihilation	positron tomography	particle and antiparticle, e.g. electron and positron, disappear into photons
Penning effect	Ne, Ar	ionization through collision with metastable atoms
Richardson effect, thermal emission	BaO_2, W, Mo, used in tv and electron microscopes	emission of electrons from hot metals
skin effect	Cu	high current density on exterior of wire
pinch effect	InSb, plasmas	high current density on interior of wire
Josephson effect	Nb-Oxide-Nb	tunnel current flows through insulator between two superconductors
Sasaki–Shibuya effect	n-Ge, n-Si	anisotropy of conductivity due to applied electric field
switchable magnetism	InAs:Mn	voltage switchable magnetization Ref. 562
Interactions with magnetic fields		
Hall effect	silicon; used for magnetic field measurements	voltage perpendicular to current flow in applied magnetic field
Zeeman effect	Cd	change of emission frequency with magnetic field
Paschen–Back effect	atomic gases	change of emission frequency in strong magnetic fields
ferromagnetism	Fe, Ni, Co, Gd	spontaneous magnetization; material strongly attracted by magnetic fields

Property	Example	Definition
paramagnetism	Fe, Al, Mg, Mn, Cr	induced magnetization parallel to applied field; attracted by magnetic fields
diamagnetism	water, Au, graphite, NaCl	induced magnetization opposite to applied field; repelled by magnetic fields
magnetostriction	CeB_6, $CePd_2Al_3$	change of shape or volume by applied magnetic field
magnetoelastic effect	Fe, Ni	change of magnetization by tension or pressure
acoustomagnetic effect	metal alloys, anti-theft etiquettes	excitation of mechanical oscillations through magnetic field
spin valve effect	metal multilayers	electrical resistance depends on spin direction of electrons with respect to applied magnetic field
magnetooptical activity or Faraday effect or Faraday rotation	flint glass	polarization angle is rotated with magnetic field; different refraction index for right and left circularly polarized light, as in magnetooptic (MO) recording
magnetic circular dichroism	gases	different absorption for right and left circularly polarized light; essentially the same as the previous one
Majorana effect	colloids	specific magnetooptic effect
photoelectromagnetic effect	InSb	current flow due to light irradiation of semiconductor in a magnetic field
Voigt effect	vapours	birefringence induced by applied magnetic field
Cotton–Mouton effect	liquids	birefringence induced by applied magnetic field
Hanle effect	Hg	change of polarization of fluorescence with magnetic field
Shubnikov–de Haas effect	Bi	periodic change of resistance with applied magnetic field
thermomagnetic effects: Ettinghausen effect, Righi–Leduc effect, Nernst effect, magneto–Seebeck effect	BiSb alloys	relation between temperature, applied fields and electric current
Ettinghausen–Nernst effect	Bi	appearance of electric field in materials with temperature gradients in magnetic fields
photonic Hall effect	CeF_3	transverse light intensity depends on the applied magnetic field Ref. 563
magnetocaloric effect	gadolinium, GdSiGe alloys	material cools when magnetic field is switched off Ref. 564

Property	Example	Definition
cyclotron resonance	semiconductors, metals	selective absorption of radio waves in magnetic fields
magnetoacoustic effect	semiconductors, metals	selective absorption of sound waves in magnetic fields
magnetic resonance	most materials, used for imaging in medicine for structure determination of molecules	selective absorption of radio waves in magnetic fields
magnetorheologic effect	liquids, used in advanced car suspensions	change of viscosity with applied magnetic fields
Meissner effect	type 1 superconductors, used for levitation	expulsion of magnetic field from superconductors
Interactions with electric fields		
polarizability	all matter	polarization changes with applied electric field
ionization, field emission, Schottky effect	all matter, tv	charges are extracted at high fields
paraelectricity	$BaTiO_3$	applied field leads to polarization in same direction
dielectricity	water	in opposite direction
ferroelectricity	$BaTiO_3$	spontaneous polarization below critical temperature
piezoelectricity	like the quartz lighter used in the kitchen	polarization appears with tension, stress, or pressure
electrostriction	platinum sponges in acids	shape change with applied voltage Ref. 565
pyroelectricity	$CsNO_3$, tourmaline, crystals with polar axes; used for infrared detection	change of temperature produces charge separation
electroosmosis or electrokinetic effect	many ionic liquids	liquid moves under applied electric field Ref. 566
electrowetting	salt solutions on gold	wetting of surface depends on applied voltage
electrolytic activity	sulfuric acid	charge transport through liquid
liquid crystal effect	watch displays	molecules turn with applied electric field
electrooptical activity: Kerr effect, Pockels effect	liquids (e.g. oil), crystalline solids	material in electric field rotates light polarization, i.e. produces birefringence
Freederichsz effect, Schadt–Helfrichs effect	nematic liquid crystals	electrically induced birefringence

Property	Example	Definition
Stark effect	hydrogen, mercury	colour change of emitted light in electric field
field ionization	helium near tungsten tips in field ion microscope	ionization of gas atoms in strong electric fields
Zener effect	Si	energy-free transfer of electrons into conduction band at high fields
field evaporation	W	evaporation under strong applied electric fields
Interactions with light		
absorption	coal, graphite	transformation of light into heat or other energy forms (which ones?)Challenge 1045 n
blackness	coal, graphite	complete absorption in visible range
colour, metallic shine	ruby	absorption depending on light frequency
photostriction	PbLaZrTi	light induced piezoelectricity
photography	AgBr, AgI	light precipitates metallic silver
photoelectricity, photoeffect	Cs	current flows into vacuum due to light irradiation
internal photoelectric effect	Si p–n junctions, solar cells	voltage generation and current flow due to light irradiation
photon drag effect	p-Ge	current induced by photon momentum
emissivity	every body	ability to emit light
transparency	glass, quartz, diamond	low reflection, low absorption, low scattering
reflectivity	metals	light bounces on surface
polarization	pulled polymer sheets	light transmission depending on polarization angle
optical activity	sugar dissolved in water, quartz	rotation of polarization
birefringence	feldspar,cornea	refraction index depends on polarization direction, light beams are split into two beams
dichroism	feldspar, andalusite	absorption depends on polarization
optically induced anisotropy, Weigert effect	AgCl	optically induced birefringence and dichroism
second harmonic generation	$LiNbO_3$, KPO_4	light partially transformed to double frequency
luminescence: general term for opposite of incandescence	GaAs, television	cold light emission
fluorescence	CaF_2, X ray production, light tubes, cathode ray tubes	light emission during and after light absorption or other energy input

Property	Example	Definition
phosphorescence	$TbCl_3$	light emission due to light, electrical or chemical energy input, continuing *long after* stimulation
electroluminescence	ZnS	emission of light due to alternating electrical field
photoluminescence	ZnS : Cu, $SrAlO_4$: Eu, Dy, hyamine	light emission triggered by UV light, used in safety signs
chemoluminescence	H_2O_2, phenyl oxalate ester, dye	cold light emission used in light sticks for divers and fun
bioluminescence	glow-worm, deep sea fish	cold light emission in animals
triboluminescence	sugar	light emission during friction or crushing
thermoluminescence	quartz, feldspar	light emission during heating, used e.g. for archaeological dating of pottery Ref. 567
Bremsstrahlung	X ray generation	radiation emission through fast deceleration of electrons
Compton effect	momentum measurements	change of wavelength of light, esp. X rays and gamma radiation, colliding with matter
Čerenkov effect	water, polymer particle detectors	light emission in a medium due to particles, e.g. emitted by radioactive processes, moving faster than the speed of light in that medium
transition radiation	any material	light emission due to fast particles moving from one medium to a second with different refractive index
electrochromicity	wolframates	colour change with applied electric field
scattering	gases, liquids	light changes direction
Mie scattering	dust in gases	light changes direction
Raleigh scattering	sky	light changes direction, sky is blue
Raman effect or Smekal–Raman effect	molecular gases	scattered light changes frequency
laser activity, superradiation	beer, ruby, He–Ne	emission of stimulated radiation
sonoluminescence	air in water	light emission during cavitation
gravitoluminescence	fake; it does not exist; why? Challenge 1046 n	
switchable mirror	LaH	voltage controlled change from reflection to transparency Ref. 568
radiometer effect	bi-coloured windmills	mill turn due to irradiation (see page 529)
luminous pressure	idem	opposite of the preceding one
solar sail effect	future satellites	motion due to solar wind

Property	Example	Definition
acoustooptic effect	$LiNbO_3$	diffraction of light by sound in transparent materials
photorefractive materials	$LiNbO_3$, GaAs, InP	light irradiation changes refractive index
Auger effect	Auger electron spectroscopy	electron emission due to atomic reorganization after ionization by X rays
Bragg reflection	crystal structure determination	X ray diffraction by atomic planes
Mößbauer effect	Fe, used for spectroscopy	recoil-free resonant absorption of gamma radiation
pair creation	Pb	transformation of a photon in a charged particle–antiparticle pair
photoconductivity	Se, CdS	change of resistivity with light irradiation
optoacoustic affect, photoacoustic effect	gases, solids	creation of sound due to absorption of pulsed light
optogalvanic effect	plasmas	change of discharge current due to light irradiation
optical nonlinear effects: parametric amplification, frequency mixing, saturable absorption, n-th harmonic generation, optical Kerr effect, etc.		
phase conjugated mirror activity	gases	reflection of light with opposite phase
Material properties		
solidity, impenetrability	floors, columns, ropes, buckets	at most one object per place at a given time
Interactions with vacuum		
Casimir effect	metals	attraction of uncharged, conducting bodies

All matter properties in the list can be influenced by electric or magnetic fields or directly depend on them. This shows that the nature of all these material properties is electromagnetic. In other words, charges and their interactions are an essential and fundamental part of the structure of objects. The table shows so many different electromagnetic properties that the motion of charges inside each material must be complex indeed. Most effects are the topic of solid state physics,* fluid and plasma physics.

Solid state physics is by far the most important part of physics, when measured by the impact it had on society. Almost all effects have applications in technical products, and give work to many people. Can you name a product or business application for any randomly chosen effect from the table? Challenge 1047 e

In our mountain ascent however, we look only at one example from the above list: thermal radiation, the emission of light by hot bodies.

* Probably the best and surely the most entertaining introductory English language book on the topic is the one by Neil Ashcroft & David Mermin, *Solid State Physics*, Holt Rinehart & Winston, 1976.

Earnshaw's theorem about the impossibility of a stable equilibrium for charged particles at rest implies that the charges inside matter must be *moving*. For any charged particle in motion, Maxwell's equations for the electromagnetic field show that it radiates energy by emitting electromagnetic waves. In short, classical mechanics thus predicts that matter must radiate electromagnetic energy.

Interestingly, everybody knows from experience that this is indeed the case. Hot bodies light up depending on their temperature; the working of light bulbs thus proves that metals are made of charged particles. *Incandescence*, as it is called, requires charges. Actually, *every* body emits radiation, even at room temperature. This radiation is called *thermal radiation*; at room temperature it lies in the infrared. Its intensity is rather weak
Ref. 569 in everyday life; it is given by the general expression

$$I(T) = f\,T^4\,\frac{2\pi^5 k^4}{15c^2h^3} \quad \text{or} \quad I(T) = f\sigma T^4 \quad \text{with} \quad \sigma = 56.7\,\mathrm{nW/K^4m^2} \qquad (454)$$

where f is a material, shape and temperature dependent factor, with a value between zero and one, and called the *emissivity*. The constant σ is also called the *Stefan–Boltzmann black body radiation constant* or *black body radiation constant*. A body whose emissivity is given by the ideal case $f = 1$ is called a *black body*, because at room temperature such
Challenge 1048 n bodies also have an ideal absorption coefficient and thus appear black. (Can you see why?) The heat radiation they emit is called *black body radiation*.

Ref. 570 By the way, which object radiates more energy: a human body or an average piece of
Challenge 1049 n the Sun of the same mass? Guess first!

Why can we see each other?

Physicists have a strange use of the term 'black'. Most bodies at temperatures at which they are red hot or even hotter are excellent approximations of black bodies. For example, the tungsten in incandescent light bulbs, at around 2000 K, emits almost pure black body radiation; however, the glass then absorbs much of the ultraviolet and infrared components. Black bodies are also used to define the colour *white*. What we commonly call pure white is the colour emitted by a black body of 6500 K, namely the Sun. This definition is used
Ref. 571 throughout the world, e.g. by the Commission Internationale d'Eclairage. Hotter black bodies are bluish, colder ones are yellow, orange or red.* The stars in the sky are classified in this way, as summarized on page 170.

Let us have a quick summary of black body radiation. Black body radiation has two important properties; first, the emitted light power increases with the fourth power of the temperature. With this relation alone you can check the just mentioned temperature of the Sun simply by comparing the size of the Sun with the width of your thumb when
Challenge 1050 d the arm is stretched away from the face. Are you able to do this? (Hint: use the excellent
Ref. 572 approximation that the Earth's average temperature of about 14.0°C is due to the Sun's irradiation.)

* Most bodies are not black, because colour is not only determined by emission, but also by absorption of light.

The precise expression for the emitted energy density u per frequency ν can be deduced from the radiation law for black bodies discovered by Max Planck in 1899:

$$u(\nu, T) = \frac{8\pi h}{c^3} \frac{\nu^3}{e^{h\nu/kT} - 1} . \tag{455}$$

He made this important discovery, which we will discuss in more detail in the second part of our mountain ascent, simply by comparing this curve with experiment.* The new constant h, *quantum of action* or *Planck's constant*, turns out to have the value $6.6 \cdot 10^{-34}$ Js, and is central to all quantum theory, as we will see. The other constant Planck introduced, the Boltzmann constant k, appears as a prefactor of temperature all over thermodynamics, as it acts as a conversion unit from temperature to energy.

Page 657

Challenge 1051 ny

The radiation law gives for the total emitted energy density the expression

$$u(T) = T^4 \frac{8\pi^5 k^4}{15 c^3 h^3} \tag{456}$$

Challenge 1052 ny

from which equation (454) is deduced using $I = uc/4$. (Why?)

The second property of black body radiation is the value of the peak wavelength, i.e. the wavelength emitted with the highest intensity. This wavelength determines their colour; it is deduced from equation (455) to be

Challenge 1053 ny

$$\lambda_{\max} = \frac{hc}{4.956\,k} \frac{1}{T} = 2.9\,\mathrm{mm\,K}/T \quad \text{but} \quad \hbar\nu_{\max} = 2.82\,kT = 3.9 \cdot 10^{-23}\,\mathrm{J/K} \cdot T . \tag{457}$$

Either of these expressions is called *Wien's colour displacement* after its discoverer.** The colour change with temperature is used in optical thermometers; that is also the way the temperature of stars is measured. For 37°C, human body temperature, it gives a peak wavelength of 9.3 µm or 115 THz, which is thus the colour of the bulk of the radiation emitted by every human being. (The peak wavelength does not correspond to the peak frequency. Why?) On the other hand, following the telecommunication laws of many countries, any radiation emitter needs a licence to operate; as a consequence in Germany only dead people are legal, and only if their bodies are at absolute zero temperature.

Challenge 1054 ny

Note that a black body or a star can be blue, white, yellow, orange or red. It is never green. Can you explain why?

Challenge 1055 ny

Above, we predicted that any material made of charges emits radiation. Are you able to find a simple argument showing whether heat radiation is or is not this classically

* Max Planck (1858–1947), professor of physics in Berlin, was a central figure in thermostatics. He discovered and named *Boltzmann's constant k* and the *quantum of action h*, often called Planck's constant. His introduction of the quantum hypothesis was the birth date of quantum theory. He also made the works of Einstein known in the physical community, and later organized a job for him in Berlin. He received the Nobel prize for physics in 1918. He was an important figure in the German scientific establishment; he also was one of the very few who had the courage to tell Adolf Hitler *face to face* that it was a bad idea to fire Jewish professors. (He got an outburst of anger as answer.) Famously modest, with many tragedies in his personal life, he was esteemed by everybody who knew him.

** Wilhelm Wien (b. 1864 Gaffken, d. 1824 München), East-Prussian physicist; he received the Nobel prize for physics in 1911 for the discovery of this relation.

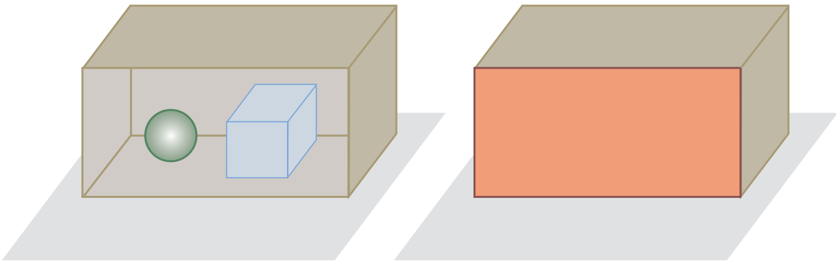


Figure 269 Bodies inside a oven at room (left) and red hot (right) temperature

Challenge 1056 ny predicted radiation?

But let us come back to the question in the section title. The existence of thermal radiation implies that any hot body will cool, even if it is left in the most insulating medium there is, namely in vacuum. More precisely, if the vacuum is surrounded by a wall, a body in the vacuum will gradually approach the same temperature as the wall.

Interestingly, when the temperature of the wall and of the body inside have become the same, something strange happens. The effect is difficult to check at home, but impressive photographs exist in the literature. Ref. 573

An arrangement in which the walls and the objects inside are at the same temperature is called an *oven*. It turns out that it is *impossible* to see objects in an oven using the light coming from thermal radiation. For example, if an oven and all its contents are red hot, taking a picture of the inside of the oven (without a flash!) does not reveal anything; no contrast nor brightness changes exist which allow to distinguish the objects from the walls or their surroundings. Can you explain the finding? Challenge 1057 n

In short, we are able to see each other only because the light sources we use are at a *different* temperature than ourselves. We can see each other only because we do *not* live in thermal equilibrium with our environment.

A summary of classical electrodynamics and of its limits

In general, classical electrodynamics can be summarized in a few main ideas.

- The electromagnetic field is a physical observable, as shown e.g. by compass needles.
- The field sources are the (moving) charges and the field evolution is described by Maxwell's evolution equations, as shown e.g. by the properties of amber, lodestone, batteries and remote controls.
- The electromagnetic field changes the motion of electrically charged objects via the Lorentz expression, as e.g. shown by electric motors.
- The field behaves like a continuous quantity, a distribution of little arrows, and propagates as a wave, as shown e.g. by radios and mobile phones.
- The field can exist and move also in empty space, as shown e.g. by the stars.

As usual, the motion of the sources and the field is reversible, continuous, conserved and deterministic. However, there is quite some fun waiting; even though this description is correct in everyday life, during the rest of our mountain ascent we will find that *each* of the bullet points is in fact wrong. A simple example shows this.

At a temperature of zero Kelvin, when matter does not radiate thermally, we have the paradoxical situation that the charges inside matter cannot be moving, since no emitted radiation is observed, but they cannot be at rest either, due to Earnshaw's theorem. In short, the simple existence of matter – with its discrete charge values – shows that classical electrodynamics is wrong.

Page 558

In fact, the overview of material properties of Table 48 makes the same point even more strongly; classical electrodynamics can describe many of the effects listed, *but it cannot explain the origin of any of them*. Even though few of the effects will be studied in our walk – they are not essential for our adventure – the general concepts necessary for their description will be the topic of the second part of this mountain ascent, that on quantum theory.

18. Classical physics in a nutshell – one and a half steps out of three

The description of general relativity and classical electrodynamics concludes our walk through classical physics. In order to see its limitations, we summarize what we have found out. In nature, we learned to distinguish and to characterize objects, radiation and space-time. All of the three can move. In all motion we distinguish the fixed, intrinsic properties from the varying state. All motion happens in a way that change is minimized.

Looking for all the *fixed, intrinsic* aspects of objects, we find that all sufficiently small objects or particles are described completely by their mass and their electric charge. There is no magnetic charge. Mass and electric charge are thus the only localized intrinsic properties of classical, everyday objects. Both mass and electric charge are defined by the accelerations they produce around them. Both quantities are conserved; thus they can be added. Mass, in contrast to charge, is always positive. Mass describes the interaction of objects with their environment, charge the interaction with radiation.

All *varying* aspects of objects, i.e. their state, can be described using momentum and position, as well as angular momentum and orientation. All can vary continuously in amount and direction. Therefore set of all possible states forms a space, the so-called *phase space*. The state of extended objects is given by the states of all its constituent particles. These particles make up all objects and somehow interact electromagnetically.

The state of a particle depends on the observer. The state is useful to calculate the change that occurs in motion. For a given particle, the change is independent of the observer, but the states are not. The states found by different observers are related: the relations are called the 'laws' of motion. For example, for different times they are called *evolution equations*, for different places and orientations they are called *transformation relations*, and for different gauges they are called *gauge transformations*. All can be condensed in the principle of least action.

We also observe motion of a massless entity: *radiation*. Everyday types of radiation, such as light, radio waves and their related forms, are travelling electromagnetic waves. They are described by same equations that describe the interaction of charged or magnetic objects. The speed of massless entities is the maximum possible speed in nature and is the same for all observers. The *intrinsic properties* of radiation are its dispersion relation and its energy–angular momentum relation. The *state* of radiation is described by its electromagnetic field strength, its phase, its polarization and its coupling to matter. The motion of radiation describes the motion of images.

The space-time *environment* is described by space and time coordinates. Space-time is also able to move, by changing its curvature. The intrinsic properties of space-time are the number of dimensions, its signature and its topology. The state is given by the metric, which describes distances and thus the local warpedness. The warpedness can oscillate and propagate, so that empty space can move like a wave.

Our environment is finite in age. It has a long history, and on large scales, all matter in the universe moves away from all other matter. The large scale topology of our environment is unclear, as is unclear what happens at its spatial and temporal limits.

Motion follows a simple rule: change is always as small as possible. This applies to matter, radiation and space-time. All energy moves in the way space-time tells it, and space moves the way energy tells it. This relation describes the motion of the stars, of thrown stones, of light beams and of the tides. Rest and free fall are the same, and gravity is curved space-time. Mass breaks conformal symmetry and thus distinguishes space from time.

Energy and mass speed is bound from above by a universal constant c, and energy change per time is bound from above by a universal constant $c^5/4G$. The speed value c is realized for motion of massless particles. It also relates space to time. The power value $c^5/4G$ is realized by horizons. They are found around black holes and at the border of the universe. The value also relates space-time curvature to energy flow and thus describes the elasticity of space-time.

No two objects can be at the same spot at the same time. This is the first statement about electromagnetism humans encounter. More detailed investigation shows that electric charge accelerates other charges, that charge is necessary to define length and time intervals, and that charges are the source of electromagnetic fields. Also light is such a field. Light travels at the maximum possible velocity. In contrast to objects, light can interpenetrate. In summary, we learned that of the two naive types of object motion, namely motion due to gravity – or space-time curvature – and motion due to the electromagnetic field, only the latter is genuine.

Above all, classical physics showed us that motion, be it linear or rotational, be it that of matter, radiation or space-time, is conserved. Motion is continuous. More than that, motion is similar to a continuous substance: it is never destroyed, never created, but always redistributed. Due to conservation, all motion, that of objects, images and empty space, is predictable and reversible. Due to conservation of motion, time and space can be defined. In addition, we found that classical motion is also right–left symmetric. Classical physics showed us that motion is predictable: there are *no* surprises in nature.

The future of planet Earth

Maybe nature shows no surprises, but it still provides many adventures. On the 8th of March 2002, a 100 m sized body almost hit the Earth. It passed at a distance of only 450 000 km. On impact, it would have destroyed a region of the size of London. A few months earlier, a 300 m sized body missed the Earth by 800 000 km; the record so far was in 1994, when the distance was only 100 000 km.* Several other adventures can be predicted by classical physics, as shown in Table 49. Many are problems facing humanity only in a distant future, but some, such as volcanic eruptions or asteroid impacts, could happen any time. All are research topics.

Ref. 576

Table 49 Examples of disastrous motion of possible future importance

Critical situation	Years from now
End of fundamental physics	*c.* 30 (around year 2030)
Giant tsunami from volcanic eruption at Canary islands	*c.* 10-200
Major nuclear material accident or weapon use	unknown
Ozone shield reduction	*c.* 100
Rising ocean levels due to greenhouse warming	*c.* 100-1 000
End of applied physics	> 200
Explosion of volcano in Greenland, leading to long darkening of sky	unknown
Several magnetic north and south poles appear, allowing solar storms to disturb radio and telecommunications, to interrupt electricity supplies, to increase animal mutations and to disorient migrating animals such as wales, birds and tortoises	*c.* 800
Our interstellar gas cloud detaches from the solar systems, changing the size of the heliosphere, and thus expose us more to aurorae and solar magnetic fields	*c.* 3 000
Reversal of Earth's magnetic field, implying a time with almost no magnetic field, with increased cosmic radiation levels and thus more skin cancers and miscarriages	unknown
Atmospheric oxygen depletion due to forest reduction and exaggerated fuel consumption	> 1000
Upcoming ice age	*c.* 15 000
Possible collision with interstellar gas cloud assumed to be crossed by the Earth every 60 million years, maybe causing mass extinctions	*c.* 50 000
Explosion of Yellowstone or other giant volcano leading to year-long volcanic winter	0 to 100 000
Possible genetic degeneration of homo sapiens due to Y chromosome reduction	*c.* 200 000
Africa collides with Europe, transforming the Mediterranean into a lake that starts evaporating	around $3 \cdot 10^6$
Gamma ray burst from within our own galaxy, causing radiation damage to many living beings	between 0 and $5 \cdot 10^6$
Asteroid hitting the Earth, generating tsunamis, storms, darkening sunlight, etc.	between 0 and $50 \cdot 10^6$
Neighbouring star approaching, starting comet shower through destabilization of Oort cloud and thus risk for life on Earth	$> 10^6$

* The web pages around http://cfa-www.harvard.edu/iau/lists/Closest.html provide more information on such events.

Critical situation	Years from now
American continent collides with Asia	$> 100 \cdot 10^6$
Instability of solar system	$> 100 \cdot 10^6$
Low atmospheric CO_2 content stops photosynthesis	$> 100 \cdot 10^6$
Collision with star cluster or other galaxy	$> 150 \cdot 10^6$
Sun ages and gets hotter, evaporating seas	$250 \cdot 10^6$
Ocean level increase due to Earth rotation slowing/stop (if not evaporated before)	$> 10^9$
Temperature rise/fall (depending on location) due to Earth rotation stop	$> 10^9$
Sun runs out of fuel, becomes red giant, engulfs Earth	$5.0 \cdot 10^9$
Sun stops burning, becomes white dwarf	$5.2 \cdot 10^9$
Earth core solidifies, removing magnetic field and thus Earth's cosmic radiation shield	$10.0 \cdot 10^9$
Nearby nova (e.g. Betelgeuse) bathes Earth in annihilation radiation	unknown
Nearby supernova (e.g. Eta Carinae) blasts over solar system	unknown
Galaxy centre destabilizes rest of galaxy	unknown
Universe recollapses – if ever (see page 344)	$> 20 \cdot 10^9$
Matter decays into radiation – if ever (see Appendix C)	$> 10^{33}$
Problems with naked singularities	unknown, controversial
Vacuum becomes unstable	unknown, controversial

Despite the fascination of the predictions, we leave aside these literally tremendous issues and continue in our adventure.

The essence of classical physics: the infinitely small implies the lack of surprises

We can summarize classical physics with a simple statement: nature lacks surprises because *classical physics is the description of motion using the concept of the infinitely small.* All concepts used so far, be they for motion, space, time or observables, assume that the infinitely small exists. Special relativity, despite the speed limit, still allows infinitely small velocities; general relativity, despite its black hole limit, still allows infinitely small force and power values. Similarly, in the description of electrodynamics and gravitation, both integrals and derivatives are abbreviations of mathematical processes that use infinitely small intermediate steps.

In other words, the classical description of nature introduces the infinitely small in the description of motion. The classical description then discovers that there are no surprises in motion. The detailed study of this question lead us to a simple conclusion: the infinitely small implies determinism.* Surprises contradict the existence of the infinitely small.

* No surprises also imply no miracles. Classical physics is thus in opposition to many religions. Indeed, many religions argue that infinity is the necessary ingredient to perform miracles. Classical physics shows that this is not the case.

On the other hand, both special and general relativity have eliminated the existence of the infinitely large. There is no infinitely large force, power, size, age or speed.

Why have we not yet reached the top of the mountain?

> The more important fundamental laws and facts of physical science have all been discovered, and these are now so firmly established that the possibility of their ever being supplanted in consequence of new discoveries is exceedingly remote... Our future discoveries must be looked for in the sixth place of decimals.
>
> Albert Michelson, 1894.*

We might think that we know nature now, like Albert Michelson did at the end of the nineteenth century. He claimed that electrodynamics and Galilean physics implied that the major laws of physics were well known. The statement is often quoted as example of flawed predictions, since it reflect an incredible mental closure to the world around them. General relativity was still unknown, and so was quantum theory.

At the end of the nineteenth century, the progress in technology due to the use of electricity, chemistry and vacuum technology had allowed to build better and better machines and apparatuses. All were built with classical physics in mind. In the years between 1890 and 1920, these classical machines completely destroyed the foundations of classical physics. Experiments with these apparatuses showed that matter is made of atoms, that electrical charge comes in smallest amounts and that nature behaves randomly. Nature does show surprises – through in a restricted sense, as we will see. Like the British empire, the reign of classical physics collapsed. Speaking simply, classical physics does not describe nature at small scales.

But even without machines, the two victorian physicists could have predicted the situation. (In fact, many more progressive minds did so.) They had overlooked a contradiction between electrodynamics and nature for which they had no excuse. In our walk so far we found that clocks and meter bars are necessarily made of matter and based on electromagnetism. But as we just saw, classical electrodynamics does not explain the stability of matter. Matter is made of small particles, but the relation between these particles, electricity and the smallest charges is not clear. If we do not understand matter, we do not yet understand space and time, since they are defined using measurement devices made of matter.

Worse, the two victorian physicists overlooked a simple fact: the classical description of nature does not allow to understand *life*. The abilities of living beings – growing, seeing, hearing, feeling, thinking, being healthy or sick, reproducing and dying – are all unexplained by classical physics. In fact, all these abilities contradict classical physics. Understanding matter and its interactions, including life itself, is therefore the aim of the second part of our ascent of Motion Mountain. The understanding will take place at small scales; to understand nature, we need to study particles. Indeed, the atomic structure of matter, the existence of a smallest charge and the existence of a smallest entropy makes

* From his address at the dedication ceremony for the Ryerson Physical Laboratory at the University of Chicago.

us question the existence of the infinitely small. There is something to explore. Doing so will lead us from surprise to surprise. To be well prepared, we first take a break.

Bibliography

459 Julian Schwinger, L.L. DeRaad, K.A. Milton & W.Y. Tsai, *Classical Electrodynamics*, Perseus, Cambridge, 1998. An excellent text on the topic by one of its greatest masters.
See also the beautiful problem book by André Butoli & Jean-Marc Lévy–Leblond, *La physique en questions – électricité et magnétisme*, Vuibert, Paris, 1999. Cited on page 506.

460 K. Hagiwara & al., *Physical Review D* **66**, p. 010001, 2002, or H.V. Klapdor-Kleingrothaus & K. Zuber, *Particle Astrophysics*, The Institute of Physics, UK, 1999. See also H. Jeon & M. Longo, Search for magnetic monopoles trapped in matter, *Physical Review Letters* **75**, pp. 1443–1447, 1995. See also A.S. Goldhaber & W.P. Trower, Resource letter MM-1: magnetic monopoles, *American Journal of Physics* **58**, pp. 429–439, 1990. Cited on page 480.

461 R. Edwards, Filling station fires spark cars' recall, New Scientist, pp. 4–5, 4 March 1995. Cited on page 480.

462 S. Desmet, F. Orban & F. Grandjean, On the Kelvin electrostatic generator, *European Journal of Physics* **10**, pp. 118–122, 1989. You can find construction plans for it also in various places on the internet. Cited on page 481.

463 For more details on various electromagnetic units, see the standard text by J.D. Jackson, *Classical Electrodynamics*, 3rd edition, Wiley, 1998. Cited on page 486.

464 See the old but beautiful papers by Richard C. Tolman & T. Dale Stewart, The electromotive force produced by the acceleration of metals, *Physical Review* **8**, pp. 97–116, 1916, Richard C. Tolman & T. Dale Stewart, The mass of the electric carrier in copper, solver and aluminium, *Physical Review* **9**, pp. 164–167, 1917, and the later but much more precise experiment by C.F. Kettering & G.G. Scott, Inertia of the carrier of electricity in copper and aluminum, *Physical Review* **66**, pp. 257–267, 1944. (Obviously the American language dropped the 'i' from aluminium during that period.) The first of these papers is also a review of the preceding attempts, and explains the experiment in detail. The last paper shows what had to taken into consideration to achieve sufficient precision. Cited on page 487.

465 This effect has first been measured by S.J. Barnett, A new electron-inertia effect and the determination of m,e for the free electron in copper/ *Philosophical Magazine* **12**, p. 349, 1931. Cited on page 487.

466 See for example C. Schiller, A.A. Koomans, T.L. van Rooy, C. Schönenberger & H.B. Elswijk, Decapitation of tungsten field emitter tips during sputter sharpening, *Surface Science Letters* **339**, pp. L925–L930, 1996. Cited on page 487.

467 L.I. Schiff & M.V. Barnhill, Gravitational-induced electric field near a metal, *Physical Review* **151**, pp. 1067–1071, 1966. F.C. Witteborn & W.M. Fairbank, Experimental

comparison of the gravitational force on freely falling electrons and metallic electrons, *Physical Review Letters* **19**, pp. 1049–1052, 1967. Cited on page 488.

468 Pierre de Maricourt, *Tractatus de magnete*, 1269. Cited on page 490.

469 The ratio of angular L to magnetic M moment is

$$\frac{L}{M} = \frac{2m}{e} \cdot \frac{1}{g} , \tag{458}$$

where e is the electron charge and m its mass. Both L and M are measurable. The first measurements were published with a g-value of 1, most probably because the authors expected the value. In later experiments, de Haas found other values. Measurements by other researchers gave values nearer to 2 than to 1, a fact that was only understood with the discovery of spin. The original publications are A. Einstein & W.J. de Haas, Proefondervinderlijk bewijs voor het bestaan der moleculaire stroomen van Ampère, *Konninklijke Akademie der Wetenschappen te Amsterdam, Verslagen* **23**, p. 1449, 1915, and A. Einstein & W.J. de Haas, Experimental proof of the existence of Ampère's molecular currents, *Konninklijke Akademie der Wetenschappen te Amsterdam, Proceedings* **18**, p. 696, 1916. Cited on page 492.

470 S.J. Barnett, Magnetization by rotation, *Physical Review* **6**, pp. 171–172, 1915, and S.J. Barnett, Magnetization by rotation, *Physical Review* **6**, pp. 239–270, 1915. Cited on page 493.

471 See J.D. Jackson, *Classical Electrodynamics*, 3rd edition, Wiley, 1998, or also R.F. Harrington, *Time Harmonic Electromagnetic Fields*, McGraw–Hill, New York, 1961. Cited on pages 496 and 506.

472 The http://suhep.phy.syr.edu/courses/modules/MM/Biology/biology2.html web site gives an introduction into brain physiology. Cited on page 496.

473 R.H. Tyler, S. Maus & H. Lühr, Magnetic signal due to ocean tidal flow identified in satellite observations, *Science* **299**, pp. 239–241, 2003. The movies derived from the data can be found on the http://www.tu-bs.de/institute/geophysik/spp/publikationen.html website. Cited on page 501.

474 J. Yang, F. Lu, L.W. Kostiuk & D.Y. Kwok, Electrokinetic microchannel battery by means of electrokinetic and microfluidic phenomena, *Journal of Micromechanics and Microengineering* **13**, pp. 963–970, 2003. Cited on page 503.

475 N. Salingaros, Invariants of the electromagnetic field and electromagnetic waves, *American Journal of Physics* **53**, pp. 361–363, 1985. Cited on page 496.

476 R. Wiltschko & W. Wiltschko, *Magnetic Orientation in Animals*, Springer, Berlin, 1995. Cited on page 490.

477 A.C. de la Torre, V⩽ c in 1820?, *European Journal of Physics* **20**, pp. L23–L24, March 1999. Cited on page 497.

478 Mark D. Semon & John R. Taylor, Thoughts on the magnetic vector potential, *American Journal of Physics* **64**, pp. 1361–1369, 1996. Cited on pages 507, 508, and 509.

479 Jean Sivardière, Simple derivation of magnetic vector potentials, *European Journal of Physics* **14**, pp. 251–254, 1993. Cited on page 507.

480 T.T. Wu & C.N. Yang, 1975, Concept of nonintegrable phase factors and global formulation of gauge fields, *Physical Review D* **12**, pp. 3845–3857, Cited on page 510.

481 Oleg D. Jefimenko, A relativistic paradox seemingly violating conservation of momentum law in electromagnetic systems, *European Journal of Physics* **20**, pp. 39–44, 1999.

Of course, the missing momentum goes into the electromagnetic field. Given that the electromagnetic momentum is given by the vector potential, are you able to check whether everything comes out right? Cited on page 506.

Challenge 1058 ny

482 H. Van Dam & E.P. Wigner, Classical relativistic mechanics of interacting point particles, *Physical Review* **136B**, pp. 1576–1582, 1965. Cited on page 506.

483 An electrodynamics text completely written with (mathematical) forms is Kurt Meetz & Walter L. Engl, Elektromagnetische Felder – mathematische und physikalische Grundlagen, Springer, 1980. Cited on page 510.

484 J. Travis, Twirl those organs into place – getting to the heart of how a heart knows left from right, *Science News* **156**, 21 August, 1999. A good book on asymmetries in nature is H. Brunner, *Rechts oder links*, Wiley–Vch, 1999. Cited on page 512.

485 See for example the discussion by M.C. Corballis & I.L. Beale, On telling left from right, *Scientific American* **224**, pp. 96–104, March 1971. Cited on page 513.

486 For a captivating account on the history of the ideas on light, see David Park, *The Fire Within the Eye: a Historical Essay on the Nature and Meaning of Light*, Princeton University Press, 1997. Cited on page 516.

487 J. Lepak & M. Crescimanno, Speed of light measurement using ping, electronic preprint available at http://www.arxiv.org/abs/physics/0201053. Cited on page 517.

488 See the text by Raymond L. Lee & Alistair B. Fraser, *The Rainbow Bridge: Rainbows in Art, Myth, and Science*, Pennsylvania State University Press, 2000. A chapter can be found at the http://www.nadn.navy.mil/Oceanography/RainbowBridge/Chapter_8.html web site. Cited on page 518.

489 The beautiful slit experiment was published by E.A. Montie, E.C. Cosman, G.W. 't Hooft, M.B. van der Mark & C.W.J. Beenakker, Observation of the optical analogue of quantised conductance of a point contact, *Nature* **350**, pp. 594–595, 18 April 1991, and in the longer version E.A. Montie, E.C. Cosman, G.W. 't Hooft, M.B. van der Mark & C.W.J. Beenakker, Observation of the optical analogue of the quantised conductance of a point contact, *Physica B* **175**, pp. 149–152, 1991. The result was also publicised in numerous other scientific magazines. Cited on pages 519 and 831.

490 A recent measurement of the frequency of light is presented in Th. Udem, A. Huber, B. Gross, J. Reichert, M. Prevedelli, M. Weitz & T.W. Hausch, Phase-coherent measurement of the hydrogen 1S–2S transition frequency with an optical frequency interval divider chain, *Physical Review Letters* **79**, pp. 2646–2649, 6 October 1997. Another is C. Schwob & al., Optical frequency measurement of the 2S-12D transitions in hydrogen and deuterium: Rydberg constant and Lamb shift determinations, *Physical Review Letters* **82**, pp. 4960–4963, 21 June 1999. Cited on page 519.

491 See for example Gábor Horváth, József Gál & Rüdiger Wehner, Why are water-seeking insects not attracted by mirages? The polarization pattern of mirages, *Naturwissenschaften* **83**, pp. 300–303, 1997. Cited on page 520.

492 On the birefringence of the eye, see L. Bour, Een eigenaardige speling der natuur, *nederlands tijdschrift voor natuurkunde* **67**, pp. 362–364, December 2001. In particular, a photograph of the eye using linear polarised illumination and taken through an analyzer shows a black *cross* inside the pupil. Cited on page 520.

493 The standard reference on the propagation of light is Max Born & Emil Wolf, Principles of Optics – Electromagnetic Theory of Propagation, Interference and Diffraction of Light, Pergamon Press, 6th edition, 1998. Cited on page 524.

494 An introduction into the topic of the 22° halo, the 46° halo, Sun dogs, and the many other arcs and bows which can be seen around the Sun, see the beautifully illustrated paper by R. Greenler, Lichterscheinungen, Eiskristalle und Himmelsarchäologie, *Physikalische Blätter* **54**, pp. 133–139, 1998, or his book *Rainbows, Halos, and Glories*, Cambridge University Press 1980. Cited on page 526.

495 James E. Faller & E. Joseph Wampler, The lunar laser reflector, *Scientific American* pp. 38–49, March 1970. Cited on page 526.

496 Neil Armstrong of Apollo 11, Jim Lovell of Apollo 8 and Apollo 13, and Jim Irwin of Apollo 15 extensively searched for it and then made negative statements, as told in *Science News* p. 423, 24 & 31 December 1994. From the space shuttle however, which circles only a few hundred kilometres above the Earth, the wall *can* be seen when the Sun is low enough such that the wall appears wider through its own shadow, as explained in *Science News* **149**, p. 301, 1996. Cited on page 527.

497 M. Shih, M. Segev & G. Salamo, *Physical Review Letters* **78**, pp. 2551–2554, 1997. See also the more readable paper by Mordechai Segev & George Stegeman, Self-trapping of optical beams: spatial solitons, *Physics Today* **51**, pp. 42–48, August 1998. Cited on page 527.

498 The first correct explanation of the light mill was given by Osborne Reynolds, On certain dimensional properties of matter in the gaseous state, *Royal Society Philosophical Transactions* Part 2, 1879. The best discussion is the one given on the web by Phil Gibbs, in the frequently asked question list of the usenet news group sci.physics; it is available at the http://www.desy.de/user/projects/Physics/light-mill.html website. Cited on page 529.

499 P. Lebedev, Untersuchungen über die Druckkräfte des Lichtes, *Annalen der Physik* **6**, pp. 307–458, 1901. He was also the first who understood that this effect is the basis for the change of direction of the tails of comets when they circle around the Sun. Cited on page 529.

500 P. Galajda & P. Ormos, *Applied Physics Letters* **78**, p. 249, 2001. Cited on page 529.

501 A short overview is given by Miles Padgett & Les Allen, Optical tweezers and spanners, *Physics World* pp. 35–38, September 1997. The original papers by Ashkin's group are A. Ashkin, J.M. Dziedzic, J.E. Bjorkholm & S. Chu, Observation of a gradient force optical trap for dielectric particles, *Optics Letters* **11**, p. 288, 1986, and A. Askin, J.M. Dziedzic & T. Yamane, Optical trapping and manipulation of single cells using infrared laser beams, *Nature* **330**, p. 769, 1987. A pedagogical explanation on optical spanners, together with a way to build one, can be found in D.N. Moothoo, J. Arlt, R.S. Conroy, F. Akerboom, A. Voit & K. Dholakia, Beth's experiment using optical tweezers, *American Journal of Physics* **69**, pp. 271–276, 2001, and in S.P. Smith, S.R. Bhalotra, A.L. Brody, B.L. Brown, E.K. Boyda & M. Prentiss, Inexpensive optical tweezers for undergraduate laboratories, *American Journal of Physics* **67**, pp. 26–35, 1999. Cited on page 529.

502 R. Beth, *Physical Review* **50**, p. 115, 1936. For modern measurements, see N.B. Simpson, K. Dholakia, L. Allen & M.J. Padgett, *Optics Letters* **22**, p. 52, 1997, and M.E.J. Frieze, T.A. Nieminen, N.R. Heckenberg & H. Rubinsztein-Dunlop, *Optics Letters* **23**, p. 1, 1998. Cited on page 530.

503 See the Latin text by Dietrich von Freiberg ... Cited on page 531.

504 J. Walker, Multiple rainbows from single drops of water and other liquids, *American Journal of Physics* **44**, pp. 421–433, 1976, and his How to create and observe a dozen rainbows in a single drop of water, *Scientific American* **237**, pp. 138–144, 1977. See also K. Sassen, Angular scattering and rainbow formation in pendant drops, *Journal of the Optical Society of America* **69**, pp. 1083–1089, 1979. Cited on page 531.

505 There are also other ways to see the green ray, for longer times, namely when a fata morgana appears at sunset. An explanation with colour photograph is contained in M. Vollmer, Gespiegelt in besonderen Düften ...– Oasen, Seeungeheuer und weitere Spielereien der Fata Morgana, *Physikalische Blätter* **54**, pp. 903–909, 1998. Cited on page 532.

506 This famous discovery is by Brent Berlin & Paul Kay, *Basic Color Terms: Their Universality and Evolution*, University of California Press, 1969. Their ongoing *world colour survey* is eagerly awaited. Of course there also are ongoing studies to find possible exceptions; the basic structure is solid, as shown in the conference proceedings C.L. Hardin & Luisa Maffi, *Colour Categories in Thought and Language*, Cambridge University Press, 1997. Cited on page 533.

507 For a thorough discussion of the various velocities connected to wave trains, see the classic text by Louis Brillouin, *Wave Propagation and Group Velocity*, Academic Press, New York, 1960. It expands in detail the theme discussed by Arnold Sommerfeld, Über die Fortpflanzung des Lichtes in dispergierenden Medien, *Annalen der Physik,* 4th series, **44**, pp. 177–202, 1914. See also Arnold Sommerfeld, *Optik*, Dietrichssche Verlagsbuchandlung, Wiesbaden 1950, section 22. An English translation Arnold Sommerfeld, *Lectures on Theoretical Physics: Optics*, 1954, is also available. Cited on pages 533 and 535.

508 Such an experiment was carried out by S. Chu & S. Wong, Linear pulse propagation in an absorbing medium, *Physical Review Letters* **48**, pp. 738–741, 1982. See also S. Chu & D. Styer, Answer to question #52. Group velocity and energy propagation, *American Journal of Physics* **66**, pp. 659–661, 1998. Another example was described in 1993 by the group of Raymond Chiao for the case of certain nonlinear materials in Raymond Chiao, Paul G. Kwait & Aephraim M. Steinberg, Faster than light?, *Scientific American* **269**, p. 52, August 1993, and R.Y. Chiao, A.E. Kozhekin & G. Kurizki, Tachyonlike excitations in inverted two-level media, *Physical Review Letters* **77**, pp. 1254–1257, 1996. On still another experimental setup using anomalous dispersion in caesium gas, see L.J. Wang, A. Kuzmich & A. Dogarin, Gain-assisted superluminal light propagation, *Nature* **406**, pp. 277–279, 20 July 2000. Cited on page 534.

509 Y.P. Terletskii, *Paradoxes in the Theory of Relativity*, Plenum Press, New York, 1968, Cited on page 535.

510 See the excellent explanation by Kirk T. McDonald, Negative group velocity, *American Journal of Physics* **69**, pp. 607–614, 2001. Cited on page 535.

511 S.A. Ramakrishna, Physics of negative refractive index materials, *Reorts on Progress of Physics* **68**, pp. 449–521, 2005. Cited on page 536.

512 The prediction of negative refraction is due to V.G. Veselago, The electrodynamics of substances with simultaneously negative values of ε and μ, *Soviet Physics Uspekhi* **10**, p. 509, 1968. The explanation with different refraction directions was published by P.M. Valanju, R.M. Walser & A.P. Valanju, Wave refraction in negative-index media: always positive and very inhomogeneous, *Physical Review Letters* **88**, p. 187401, 8 May 2002. Also Fermat's principle is corrected, as explained in V.G. Veselago, About the wording of Fermat's principle for light propagation in media with negative refraction index, http://www.arxiv.org/abs/cond-mat/0203451. Cited on page 536.

513 The first example of material system with a negative refraction index were presented by David Smith and his team. R.A. Schelby, D.R. Smith & S. Schultz, Experimental verification of a negative index of refraction, *Science* **292**, p. 77-79, 2001. More recent examples are A.A. Houck, J.B. Brock & I.L. Chuang, Experimental observations of a left-handed material that obeys Snell's law, *Physical Review Letters* **90**, p. 137401, 2003, C.G. Parazzoli, R.B. Greegor, K. Li, B.E.C. Koltenbah & M. Tanielian, Experimental verification and simulation of negative index of refraction using Snell's law, *Physical Review Letters* **90**, p. 107401, 2003. S. Foteinopoulou, E.N. Economou & C.M. Soukoulis, Refraction in media with a negative refractive index, *Physical Review Letters* **90**, p. 107402, 2003. Cited on page 536.

514 J. Pendry, Negegative refraction makes a perfect lens, *Physical Review Letters* **85**, p. 3966, 2000. Cited on page 536.

515 G. Nimtz, A. Enders & H. Spieker, *Journal de Physique I (Paris)* **4**, p. 565, 1994. Unfortunately, Nimtz himself seems to believe that he transported energy or signals faster than light; he is aided by the often badly prepared critics of his quite sophisticated experiments. See A. Enders & G. Nimtz, *Physikalische Blätter* **49**, p. 1119, Dezember 1993, and the weak replies in *Physikalische Blätter* **50**, p. 313, April 1994. See also A.M. Steinberg, *Journal de Physique I (Paris)* **4**, p. 1813, 1994, A. Steinberg & al., *Physical Review Letters* **71**, pp. 708–711, 1993, and A. Ranfagni & al., *Physical Review E* **48**, p. 1453, 1993. Cited on page 537.

516 A summary of all evidence about the motion of the aether is given by R.S. Shankland & al., New analysis of the interferometer observations of Dayton C. Miller, *Review of Modern Physics* **27**, p. 167, 1955. An older text is H. Witte, *Annalen der Physik* **26**, p. 235, 1908. Cited on page 538.

517 The history of the concept of vacuum can be found in the book by E. Grant, *Much Ado About Nothing*, Cambridge University Press, 1981, and in the extensive reference text by Edmund T. Whittaker, *A History of the Theories of Aether and Electricity*, Volume 1: *The Classical Theories*, Volume 2: *The Modern Theories*, Tomash Publishers, American Institute of Physics 1951, 1987. Cited on page 538.

The various *aether models* – gears, tubes, vortices – proposed in the nineteenth century were dropped for various reasons. Since many models used to explain electric and magnetic fields as motion of some entities, it was concluded that the speed of light would depend on electric or magnetic fields. One type of field was usually described by linear motion of the entities, the other by rotatory or twisting motion; both assignments are possible. As a consequence, aether must be a somewhat strange fluid which flows perfectly, but which resists rotation of volume elements, as McCullogh deduced in 1839. However, experiments show that the speed of light in vacuum does not depend on electromagnetic field intensity. Vortices were dropped because real world vortices were found out to be unstable. All models received their final blow when they failed to meet the requirements of special relativity.

518 S.W. Hell, Strategy foir far-field optical imaging and writing without diffraction limit, *Physics Letters A* **326**, pp. 140–145, 2004, see also V. Westphal & S.W. Hell, Nanoscale resolution in the focal plane of an optical microscope, *Physical Review Letters* **94**, p. 143903, 2005, and V. Westphal, J. Seeger, T. Salditt & S.W. Hell, Stimulated emission depletion microscopy on lithographic microstructures, *Journal of Physics B* **38**, pp. S695–S705, 2005. Cited on page 544.

519 A.D. Erlykin & A.W. Wolfendale, The origin of cosmic rays, *European Journal of Physics* **20**, pp. 409–418, 1999, Cited on page 546.

520 D. Singleton, Electromagnetic angular momentum and quantum mechanics, *American Journal of Physics* **66**, pp. 697–701, 1998, Cited on page 547.

521 A. Yazdani, D.M. Eigler & N.D. Lang, Off-resonance conduction through atomic wires, *Science* **272**, pp. 1921–1924, 28 June 1996. For aluminium, gold, lead, niobium, as well as the influence of chemical properties, see Elke Scheer, The signature of chemical valence in the electric conduction through a single-atom contact, *Nature* **394**, pp. 154–157, 9 July 1998. Cited on page 547.

522 See L. Kowalski, A myth about capacitors in series, *The Physics Teacher* **26**, pp. 286–287, 1988, and A.P. French, Are the textbook writers wrong about capacitors?, *The Physics Teacher* **31**, pp. 156–159, 1993. Cited on page 547.

523 This problem was suggested by Vladimir Surdin. Cited on page 548.

524 See for example, J.M. Aguirregabiria, A. Hernandez & M. Rivas, Velocity fields inside a conducting sphere near a slowly moving charge, *American Journal of Physics* **62**, pp. 462–466, 1994. Cited on page 548.

525 Philip Cohen, Open wide, this won't hurt a bit, *New Scientist* p. 5, 3 February 1996. Cited on page 548.

526 J.E. Avron, E. Berg, D. Goldsmith & A. Gordon, Is the number of photons a classical invariant?, *European Journal of Physics* **20**, pp. 153–159, 1999. Cited on page 549.

527 H. Montgomery, Unipolar induction: a neglected topic in the teaching of electromagnetism, *European Journal of Physics* **20**, pp. 271–280, 1999. Cited on page 501.

528 Such a claim was implicitly made by D. Mugnai, A. Ranfagni & R. Ruggieri, Observation of superluminal behaviors in wave propagation, *Physical Review Letters* **84**, p. 4830, 2000. An excellent explanation and rebuttal was given vy W.A. Rodrigues, D.S. Thober & A.L. Xavier, Causal explanation for observed superluminal behavior of microwave propagation in free space, http://www.arxiv.org/abs/physics/0012032. Cited on page 549.

529 If you want to see more on how the world looks for the different types of colour blind, have a look at the http://webexhibits.org/causesofcolor/2.html or the http://www.vischeck.com/examples web pages. Cited on page 550.

530 About the life-long passion that drove Luke Howard, see the book by Richard Hamblyn, *The Invention of Clouds*, Macmillan 2001. Cited on page 551.

531 J. Latham, The electrification of thunderstorms, *Quartely Journal of the Royal Meteorological Society* **107**, pp. 277–289, 1981. For a more recent and wider review, see Earle R. Williams, The tripole structure of thunderstorms, *Journal of Geophysical Research* **94**, pp. 13 151–13 167, 1989. See also the book by the National Research Council Staff, *The Earth's Electrical Environment*, Studies in Geophysics, National Academy Press, 1986. Cited on page 551.

532 B.M. Smirnov, Physics of ball lightning, *Physics Reports* **224**, pp. 151–236, 1993. See also D. Finkelstein & J. Rubinstein, Ball lightning, *Physical Review* **135**, pp. 390–396, 1964. For more folklore on the topic, just search the world wide web. Cited on page 551.

533 A.V. Gurevich & K.P. Zybin, Runaway breakdown and the mysteries of lightning, *Physics Today* **58**, pp. 37–43, May 2005. Cited on page 551.

534 The use of Schumann resonances in the Earth–ionosphere capacitor for this research field is explained in K. Schlegel & M. Füllerkrug, Weltweite Ortung von Blitzen, *Physik in unserer Zeit* **33**, pp. 257–261, 2002. Cited on page 552.

535 J.R. Dwyer, M.A. Uman, H.K. Rassoul, M. Al-Dayeh, E.L. Caraway, J. Jerauld, V.A. Rakov, D.M. Jordan, K.J. Rambo, V. Corbin & B. Wright, Energetic radiation produced by rocket-triggered lightning, *Science* **299**, pp. 694–697, 2003. Cited on page 553.

536 J.R. Dwyer, A fundamental limit on electric fields in air, *Geophysical Research Letters* **30**, p. 2055, 2003. Cited on page 553.

537 This happened to Giovanni Bellini (*c.* 1430–1516) the great venetian Renaissance painter, who even put this experience into writing, thus producing one of the greatest 'gaffes' ever. Cited on page 539.

538 S.R. Wilk, How retroreflectors really work, *Optics & Photonics News,* pp. 6–7, December 1993. Cited on page 539.

539 W.H. Ehrenstein & B. Lingelbach, Das Hermann–Gitter, *Physik in unserer Zeit* **6**, pp. 263–268, 2002. The journal also shows a colour variation of these lattices. Cited on page 540.

540 The eye sensitivity myth is debunked in detail by B.H. Soffer & D.K. Lynch, Some paradoxes, errors, and resolutions concerning the spectral optimization of human vision, *American Journal of Physics* **67**, pp. 946–953, 1999. Cited on page 541.

541 David R. Williams, Supernormal Vision, *Science News* **152**, pp. 312–313, 15 November 1997. See also http://www.cvs.rochester.edu/people/d_williams/d_williams.html as well as the photographs at http://www.cvs.rochester.edu/people/~aroorda/ao_research.html of the interior of living human eyes. Their last publication is A. Roorda, A. Metha, P. Lennie & D.R. Williams, Packing arrangement of the three cone classes in the primate retina, *Vision Research* **41**, pp. 1291–1306, 2001. Cited on page 541.

542 For a recent summary, see S. Parrott, http://www.arxiv.org/abs/gr-qc/9711027. See also T.A. Abbott & D.J. Griffiths, Acceleration without radiation, *American Journal of Physics* **53**, pp. 1203–1211, 1985. See alsoA. Kovetz & G.E. Tauber, Radiation from an accelerated charge and the principle of equivalence, *American Journal of Physics* **37**, pp. 382–385, 1969. Cited on page 553.

543 C.G. Tsagas, Magnetic tension and the geometry of the universe, *Physical Review Letters* **86**, pp. 5421–5424, 2001. An overview of the topic is C.G. Tsagas, Geometrical aspects of cosmic magnetic fields, http://www.arxiv.org/abs/gr-qc/0112077. Cited on page 555.

544 On the geodynamo status, see the article by P.H. Roberts & G.A. Glatzmaier, Geodynamo theory and simulations, *Reviews of Modern Physics* **72**, pp. 1081–1123, 2000. An older article is R. Jeanloz & B. Romanowicz, Geophysical dynamics at the center of the Earth, *Physics Today* pp. 22–27, August 1997. Cited on pages 502 and 554.

545 This is deduced from the $g-2$ measurements, as explained in his Nobel prize talk by Hans Dehmelt, Experiments with an isolated subatomic particle at rest, *Reviews of Modern Physics* **62**, pp. 525–530, 1990, and in Hans Dehmelt, Is the electron a composite particle?, *Hyperfine Interactions* **81**, pp. 1–3, 1993. Cited on page 554.

546 A good and short introduction is the paper F. Rohrlich, The self-force and radiation reaction, *American Journal of Physics* **68**, pp. 1109–1112, 2000. Cited on page 555.

547 An excellent review is E.H. Brandt, Levitation in Physics, *Science* **243**, pp. 349–355, 1989. Cited on pages 555 and 557.

548 See the article by R. Tuckermann, S. Bauerecker & B. Neidhart, Levitation in Ultraschallfeldern – Schwebende Tröpfchen, *Physik in unserer Zeit* **32**, pp. 69–75, February 2001. Liquid drops up to 1 g have been levitated in this way. Cited on page 556.

549 F.C. Moon & P.Z. Chang, *Superconducting Levitation – Applications to Bearings and Magnetic Transportation*, Wiley & Sons, 1994. Cited on pages 556 and 557.

550 W.T. Scott, Who was Earnshaw?, *American Journal of Physics* **27**, pp. 418–419, 1959. Cited on page 556.

551 The trick is to show that $\operatorname{div}\mathbf{E}=0$, $\operatorname{curl}\mathbf{E}=0$, thus $E\nabla^2 E=0$, and from this, $\nabla^2 E^2 \geqslant 0$; there are thus no local electric field maxima in the absence of free charges. The same proof works for the magnetic field. However, bodies with dielectric constants *lower* than their environment *can* be levitated in static electric fields. An example are gas bubbles in liquids, as shown by T.B. Jones & G.W. Bliss, Bubble dielectrophoresis, *Journal of Applied Physics* **48**, pp. 1412–1417, 1977. Cited on page 556.

552 However, it is possible to levitate magnets if one uses a combination containing diamagnets. See A.K. Geim, M.D. Simon, M.I. Boamfa & L.O. Heflinger, Magnet levitation at your fingertips, *Nature* **400**, pp. 323–324, 1999. Cited on page 557.

553 The first photographs of a single *ion* was in W. Neuhauser, M. Hohenstatt, P.E. Toschek & H. Dehmelt, Localised visible Ba^+ mono-ion oscillator, *Physical Review A* **22**, pp. 1137–1140, 1980. See also D.J. Wineland & W.M. Itano, *Physics Letters A* **82**, p. 75, 1981, as well as F. Dietrich & H. Walter, *Physical Review Letters* **58**, p. 203, 1987.

For single *atoms*, see photographs in Z. Hu & H.J. Kimble, *Optics Letters* **1**, p. 1888, 1994, F. Ruschewitz, D. Bettermann, J.L. Peng & W. Ertmer, *Europhysics Letters* **34**, p. 651, 1996, D. Haubrich, H. Schadwinkel, F. Strauch, B. Ueberholz, R. Wynands & D. Meschede, *Europhysics Letters* **34**, p. 663, 1996. Cited on page 557.

554 See for example Mark Buchanan, And God said...let there be levitating strawberries, flying frogs and humans that hover over Seattle, *New Scientist* pp. 42–43, 26 July 1997, or C. Wu, Floating frogs, *Science News* **152**, pp. 632–363, 6 December 1997, and Molecular Magnetism takes off, *Physics World* April 1997, page 28. The experiments by Andre Geim, Jan Kees Maan, Humberto Carmona and Peter Main were made public by P. Rodgers, *Physics World* **10**, p. 28, 1997. Some of the results can be found in M.V. Berry & A.K. Geim, Of flying frogs and levitrons, *European Journal of Physics* **18**, pp. 307–313, 1997. See also their http://www-hfml.sci.kun.nl/hfml/levitate.html website. Cited on page 557.

555 The well-known toy allows levitation without the use of any energy source and is called the 'Levitron'. It was *not* invented by Bill Hones of Fascination Toys & Gifts in Seattle, as the http://www.levitron.com website explains. The toy is discussed by Ron Edge, Levitation using only permanent magnets, *Physics Teacher* **33**, p. 252, April 1995. It is also discussed in M.V. Berry, The Levitron^{TM}: an adiabatic trap for spins, *Proceedings of the Royal Society A* **452**, pp. 1207–1220, 1996, (of Berry's phase fame) as well as by M.D. Simon, L.O. Heflinger & S.L. Ridgeway, Spin stabilised magnetic levitation, *American Journal of Physics* **65**, pp. 286–92, 1997, and by T.B. Jones, M. Washizu & R. Gans, Simple theory for the Levitron, *Journal of Applied Physics* **82**, pp. 883–889, 1997. Cited on page 557.

556 The drill trick and the building of a Levitron are described in the beautiful lecture script by Josef Zweck, *Physik im Alltag*, Skript zur Vorlesung im WS 1999/2000 der Universität Regensburg. Cited on page 558.

557 The prediction about quantised levitation is by Stephen B. Haley, Length quantization in levitation of magnetic microparticles by a mesoscopic superconducting ring, *Physical Review Letters* **74**, pp. 3261–3264, 1995. The topic is discussed in more detail in Stephen B. Haley, Magnetic levitation, suspension, and superconductivity: macroscopic and mesoscopic, *Physical Review B* **53**, p. 3506, 1996, reversed in order with Stephen B. Haley, Quantized levitation of superconducting multiple-ring systems, *Physical Review B* **53**, p. 3497, 1996, as well as Stephen B. Haley, Quantized levitation by multiply-connected superconductors, LT-21 Proceedings, in *Czechoslovak Journal of Physics* **46**, p. 2331, 1996. In 1998, there was not yet an experimental confirmation (Stephen Haley, private communication). Cited on page 558.

558 All the illusions of the flying act look as if the magician is hanging on lines, as observed by many, including the author. (Photographic flashes are forbidden, a shimmery background is set up to render the observation of the lines difficult, no ring is ever actually pulled over the magician, the aquarium in which he floats is kept open to let the fishing lines pass through, always the same partner is 'randomly' chosen from the public, etc.) Information from eyewitnesses who have actually seen the fishing lines used by David Copperfield explains the reasons for these setups. The usenet news group alt.magic.secrets, in particular Tilman Haush-

err, was central in clearing up this issue in all its details, including the name of the company which made the suspension mechanism. Cited on page 558.

559 Detailed descriptions of many of these effects can be found in the excellent overview edited by Manfred von Ardenne, Gerhard Musiol & Siegfried Reball, Effekte der Physik und ihre Anwendungen, Harri Deutsch, 1997. Cited on page 558.

560 R. Buddakian, K. Weninger, R.A. Hiller & Seth J. Putterman, Picosecond discharges and stick–slip friction at a moving meniscus of mercury in glass, *Nature* **391**, pp. 266–268, 15 January 1998. See also *Science News* **153**, p. 53, 24 January 1998. Cited on page 558.

561 Henk Swagten & Reinder Coehoorn, Magnetische tunneljuncties, *nederlands tijdschrift voor natuurkunde* **64**, pp. 279–283, November 1998. Cited on page 559.

562 H. Ohno, D. Chiba, F. Matsukura, T. Omiya, E. Abe, T. Dietl, Y. Ohno & K. Ohtani, Electric-field control of ferromagnetism, *Nature* **408**, pp. 944–946, 21-28 December 2000. Cited on page 559.

563 This effect was discovered by Geert Rikken, Bart van Tiggelen & Anja Sparenberg, Lichtverstrooiing in een magneetveld, *nederlands tijdschrift voor natuurkunde* **63**, pp. 67–70, maart 1998. Cited on page 560.

564 Vitalij Pecharsky & Karl A. Gschneidner, Giant magnetocaloric effect in Gd5(Si2Ge2), *Physical Review Letters* **78**, pp. 4494–4497, 1995, and, from the same authors, Tunable magnetic regenerator alloys with a giant magnetocaloric effect for magnetic refrigeration from ~20 to ~2990 K, *Applied Physics Letters* **70**, p. 3299, 1997. Cited on page 560.

565 J. Weissmüller, R.N. Viswanath, D. Kramer, P. Zimmer, R. Würschum & H. Gleiter, Charge–Induced Reversible Strain in a Metal, *Science* **300**, pp. 312–315, 11 April 2003. Cited on page 561.

566 A. Ajdari, Electro-osmosis on inhomogeneously charged surfaces, *Physical Review Letters* **75**, pp. 755–758, 1995. Cited on page 561.

567 M.J. Aitken, *Thermoluminescence Dating*, Academic Press, 1985. The precision of the method is far worse that C14 dating, however, as shown by H. Huppertz, *Thermolumineszenzdatierung: eine methodologische Analyse aufgrund gesicherter Befunde*, Peter Lang Verlag, 2000. Cited on page 563.

568 This effect was discovered by J.N. Huiberts, R. Griessen, J.H. Rector, R.J. Wijngarden, J.P. Dekker, D.G. de Groot & N.J. Koeman, Yttrium and lanthanum hydride films with switchable optical properties, *Nature* **380**, pp. 231–234, 1996. A good introduction is R. Griessen, Schaltbare Spiegel aus Metallhydriden, *Physikalische Blätter* **53**, pp. 1207–1209, 1997. Cited on page 563.

569 See any book on thermostatics, such as Linda Reichl, *A Modern Course in Statistical Physics*, Wiley, 2nd edition, 1998. Cited on page 565.

570 The Sun emits about $4 \cdot 10^{26}$ W from its mass of $2 \cdot 10^{30}$ kg, about 0.2 mW/kg; a person with an average mass of 75 kg emits about 100 W (you can check this in bed at night) i.e. about 500 times more. Cited on page 565.

571 See its http://www.cie.co.at/cie website. Cited on page 565.

572 P.D. Jones, M. New, D.E. Parker, S. Martin & I.G. Rigor, Surface air temperature and its changes over the past 150 years, *Reviews of Geophysics* **37**, pp. 173–199, May 1999. Cited on page 565.

573 A picture of objects in a red hot oven and at room temperature is shown in C.H. Bennett, Demons, engines and the second law, *Scientific American* **255**, pp. 108–117, November 1987. Cited on page 567.

574 Wolfgang Rindler, *Essential Relativity – Special, General, and Cosmological*, revised second edition, Springer Verlag, 1977, page 247. There is also the beautiful paper by M. Le Bellac & J.-M. Lévy-Leblond, Galilean electrodynamics, *Nuovo Cimento B* **14**, p. 217, 1973, which explains the possibilities but also the problems appearing when trying to define the theory non-relativistically. Cited on page 515.

575 L.-C. Tu, J. Luo & G.T. Gilles, The mass of the photon, *Reports on Progress of Physics* **68**, pp. 77–130, 2005. Cited on page 515.

576 Bill McGuire, *A Guide to the End of the World: Everything you Never Wanted to Know*, Oxford University Press, 2002. Cited on page 570.

The Brain, Language and the Human Condition

Alles was überhaupt gedacht werden kann, kann klar gedacht werden.*
Ludwig Wittgenstein, *Tractatus*, 4.116

In our quest for increased precision in the description of all motion around us, it is time to take a break, sit down and look back. In our walk so far, which has led through mechanics, general relativity and electrodynamics, we used several concepts without defining them. Examples are 'information', 'memory', 'measurement', 'set', 'number', 'infinity', 'existence', 'universe' and 'explanation'. Each of these is a common and important term. In this intermezzo, we take a look at these concepts and try to give some simple, but sufficiently precise definitions, keeping them as provocative and entertaining as possible. For
Challenge 1059 e example, can you explain to your parents what a concept is?

The reason for studying definitions is simple. We need the clarifications in order to get to the top of Motion Mountain. Many have lost their way because of lack of clear concepts. In this situation, physics has a special guiding role. All sciences share one result: every type of change observed in nature is a form of motion. In this sense, but in this sense *only*, physics, focusing on motion itself, forms the basis for all the other sciences. In other words, the search for the famed 'theory of everything' is an arrogant expression for the search for a *theory of motion*. Even though the knowledge of motion is basic, its precise description does not imply a description of 'everything': just try to solve a marriage problem using the Schrödinger equation to note the difference.


Ludwig Wittgenstein

Given the basic importance of motion, it is necessary that in physics all statements on observations be as precise as possible. For this reason, many thinkers have investigated physical statements with particular care, using all criteria imaginable. Physics is detailed prattle by curious people about moving things. The criteria for precision appear once we
Challenge 1060 e ask: which abilities does this prattle require? You might want to fill in the list yourself.

The abilities necessary for talking are a topic of research even today. The way that the human species acquired the ability to chat about motion is studied by evolutionary biologists. Child psychologists study how the ability develops in a single human being. Physiologists, neurologists and computer scientists are concerned with the way the brain and the senses make this possible; linguists focus on the properties of the language we use, while logicians, mathematicians and philosophers of science study the general properties of statements about nature. All these fields investigate tools that are essential for

* 'Everything that can be thought at all can be thought clearly.' This and other quotes of Ludwig Wittgenstein are from the equally short and famous *Tractatus logico-philosophicus*, written in 1918, first published in 1921; it has now been translated into many other languages.

the development of physics, the understanding of motion and the specification of the undefined concepts listed above. The fields structure this intermezzo.

Evolution

> A hen is only an egg's way of making another egg.
> Samuel Butler, *Life and Habit*, 1877.

The evolution of the human species is the result of a long story that has been told in
Ref. 577 many excellent books. A summarizing table on the history of the universe is given in the
Page 415 chapter on general relativity. The almost incredible chain of events that has lead to one's own existence includes the formation of atoms, of the galaxies, the stars, the planets, the Moon, the atmosphere, the first cells, the water animals, the land animals, the mammals,
Ref. 578 the hominids, the humans, the ancestors, the family and finally oneself.

The way the particles we are made of moved during this sequence, being blown through space, being collected on Earth, becoming organized to form people, is one of the most awe-inspiring examples of motion. Remembering this fantastic sequence of motion
Challenge 1061 e every now and then can be an enriching experience.

In particular, without biological evolution, we would not be able to talk about motion; only moving bodies can study moving bodies. Evolution was also the fount of childhood and curiosity. In this intermezzo we will discover that most concepts of classical physics have already been introduced by little children, in the experiences they have while growing up.

Children and physics

> Physicists also have a shared reality. Other than that, there isn't really a lot of difference between being a physicist and being a schizophrenic.
> Richard Bandler

During childhood, everybody is a physicist. When we follow our own memories back-
Ref. 579 wards in time as far as we can, we reach a certain stage, situated before birth, which forms the starting point of human experience. In that magic moment, we sensed somehow that apart from ourselves, there is something else. The first observation we make about the world, during the time in the womb, is thus the recognition that we can distinguish two parts: ourselves and the rest of the world. This distinction is an example – perhaps the first – of a large number of 'laws of nature' that we stumble upon in our lifetime. By discovering more and more distinctions we bring structure in the chaos of experience. We quickly find out that the world is made of related parts, such as mama, papa, milk, Earth, toys, etc.

Later, when we learn to speak, we enjoy using more difficult words and we call the surroundings the *environment*. Depending on the context, we call the whole formed by oneself and the environment together the (physical) *world*, the (physical) *universe*, *nature*, or the *cosmos*. These concepts are not distinguished from each other in this walk;* they

* The differences in usage can be deduced from their linguistic origins. 'World' is derived from old Germanic

are all taken to designate the sum of all parts and their relations. They are simply taken here to designate the *whole*.

The discovery of the first distinction starts a chain of similar discoveries. We extract the numerous distinctions that are possible in the environment, in our own body and in the various types of interactions between them. The ability to distinguish is the central ability that allows us to change our view from that of the world as *chaos*, i.e. as a big mess, to that of the world as a *system*, i.e. a structured set, in which parts are related in specific ways. (If you like precision, you may ponder whether the two choices of 'chaos' and 'system' are the only possible ones. We will return to this issue in the third part of our mountain ascent.)

Challenge 1062 n

In particular, the observation of the differences between oneself and the environment goes hand in hand with the recognition that not only are we not independent of the environment, but we are firmly tied to it in various inescapable ways: we can fall, get hurt, feel warm, cold, etc. Such relations are called *interactions*. Interactions express the observation that even though the parts of nature can be distinguished, they cannot be isolated. In other words, interactions describe the difference between the whole and the sum of its parts. No part can be defined without its relation to its environment. (Do you agree?)

Page 628
Challenge 1063 e

Interactions are not arbitrary; just take touch, smell or sight as examples. They differ in reach, strength and consequences. We call the characteristic aspects of interactions *patterns of nature*, or *properties of nature*, or *rules of nature* or, equivalently, with their historical but unfortunate name, *'laws' of nature*. The term 'law' stresses their general validity; unfortunately, it also implies design, aim, coercion and punishment for infringement. However, no design, aim or coercion is implied in the properties of nature, nor is infringement possible. The ambiguous term 'law of nature' was made popular by René Descartes (1596–1650) and has been adopted enthusiastically because it gave weight to the laws of the state – which were far from perfect at that time – and to those of other organizations – which rarely are. The expression is an anthropomorphism coined by an authoritarian world view, suggesting that nature is 'governed'. We will therefore use the term as rarely as possible in our walk and it will, if we do, be always between 'ironical' parentheses. Nature cannot be forced in any way. The 'laws' of nature are not obligations for nature or its parts, they are obligations only for physicists and all other people: the patterns of nature oblige us to use certain descriptions and to discard others. Whenever one says that 'laws govern nature' one is talking nonsense; the correct expression is *rules describe nature*.

During childhood we learn to distinguish between interactions with the environment (or *perceptions*): some are shared with others and called *observations*, others are uniquely personal and are called *sensations*.* A still stricter criterion of 'sharedness' is used to divide the world into 'reality' and 'imagination' (or 'dreams'). Our walk will show that this distinction is not essential, provided that we stay faithful to the quest for ever increasing

'wer' – person – and 'ald' – old – and originally means 'lifetime'. 'Universe' is from the Latin, and designates the one – 'unum' – which one sees turning – 'vertere', and refers to the starred sky at night which turns around the polar star. 'Nature' comes also from the Latin, and means 'what is born'. 'Cosmos' is from Greek κόσμος and originally means 'order'.

* A child that is unable to make this distinction among perceptions – and who is thus unable to lie – almost surely develops or already suffers from *autism*, as recent psychological research has shown.

Ref. 580

precision: we will find that the description of motion that we are looking for does not depend on whether the world is 'real' or 'imagined', 'personal' or 'public'.

Humans enjoy their ability to distinguish parts, which in other contexts they also call *details*, *aspects* or *entities*, and enjoy their ability to associate them or to observe the *relations* between them. Humans call this activity *classification*. Colours, shapes, objects, mother, places, people and ideas are some of the entities that humans discover first.

Ref. 581

Our anatomy provides a handy tool to make efficient use of these discoveries: *memory*. It stores a large amount of input that is called *experience* afterwards. Memory is a tool used by both young and old children to organize their world and to achieve a certain security in the chaos of life.

Memorized classifications are called *concepts*. Jean Piaget was the first researcher to describe the influence of the environment on the concepts that a child forms. Step by step, children learn that objects are localized in space, that space has three dimensions, that objects fall, that collisions produce noise, etc. In particular, Piaget showed that space and time are not a priori concepts, but result from the interactions of every child with its environment.*

Around the time that a child goes to school, it starts to understand the idea of *permanence of substances*, e.g. liquids, and the concept of *contrary*. Only at that stage does its subjective experience becomes *objective*, with abstract comprehension. Still later, the child's description of the world stops to be animistic: before this step, the Sun, a brook or a cloud are *alive*. In short, only after puberty does a human become ready for physics.

Ref. 583

Even though everyone has been a physicist in their youth, most people remain *classical* physicists. In this adventure we continue, using all the possibilities of a toy with which

* An overview of the origin of developmental psychology is given by J.H. Flavell, *The Developmental Psychology of Jean Piaget*, 1963. This work summarizes the observations by the French speaking Swiss Jean Piaget (1896–1980), the central figure in the field. He was one of the first researchers to look at child development in the same way that a physicist looks at nature: carefully observing, taking notes, making experiments, extracting hypotheses, testing them, deducing theories. His astonishingly numerous publications, based on his extensive observations, cover almost all stages of child development. His central contribution is the detailed description of the stages of development of the cognitive abilities of humans. He showed that all cognitive abilities of children, the formation of basic concepts, their way of thinking, their ability to talk, etc., result from the continuous interaction between the child and the environment.

In particular, Piaget described the way in which children first learn that they are different from the external environment, and how they then learn about the physical properties of the world. Of his many books related to physical concepts, two especially related to the topic of this walk are J. Piaget, *Les notions de mouvement et de vitesse chez l'enfant*, Presses Universitaires de France, 1972 and *Le developpement de la notion de temps chez l'enfant*, Presses Universitaires de France, 1981, this last book being born from a suggestion by Albert Einstein. These texts should be part of the reading of every physicist and science philosopher interested in these questions.

Piaget also describes how in children the mathematical and verbal intelligence derives from sensomotorial, practical intelligence, which itself stems from habits and acquired associations to construct new concepts. Practical intelligence requires the system of reflexes provided by the anatomical and morphological structure of our organism. Thus his work shows in detail that our faculty for mathematical description of the world is based, albeit indirectly, on the physical interaction of our organism with the world.

Some of his opinions on the importance of language in development are now being revised, notably through the rediscovery of the work of Lev Vigotsky, who argues that all higher mental abilities, emotions, recollective memory, rational thought, voluntary attention and self-awareness, are not innate, but learned. This learning takes place through language and culture, and in particular through the process of talking to oneself.

Ref. 582

nature provides us: the brain.

> Experience is the name everyone gives to their mistakes.
> Oscar Wilde (b. 1854 Dublin, d. 1900 Paris), *Lady Windermere's Fan*.

Why a brain?

> Denken ist bereits Plastik.*
> Joseph Beuys (1920–1986), sculptor.

Ref. 584

Numerous observations show that sense input is processed, i.e. classified, stored and retrieved in the brain. Notably, lesions of the brain can lead to the loss of part or all of these functions. (Ref. 585) Among the important consequences of these basic abilities of the brain are thought and language. All such abilities result from the construction, from the 'hardware' of the brain. (Ref. 586)

Systems with the ability to deduce classifications from the input they receive are called *classifiers*, and are said to be able to *learn*. Our brain shares this property with many complex systems; the brain of many animals, but also certain computer algorithms, such as the so-called 'neural networks', are examples of such classifiers. Such systems are studied in several fields, from biology to neurology, mathematics and computer science.** Classifiers have the double ability to discriminate and to associate; both are fundamental to thinking.

Machine classifiers have a lot in common with the brain. As an example, following an important recent hypothesis in evolutionary biology, (Ref. 587) the necessity to cool the brain in an effective way is responsible for the upright, bipedal walk of humans. The brain needs a powerful cooling system to work well. In this it resembles modern computers, which usually have powerful fans or even water cooling systems built into them. It turns out that the human species has the most powerful cooling system of all mammals. An upright posture allowed the air to cool the body most effectively in the tropical environment where humans evolved. For even better cooling, humans have also no body hair, except on their head, where it protects the brain from direct heating by the Sun.***

All classifiers are built from smallest classifying entities, sometimes large numbers of them. Usually, the smallest units can classify input into only two different groups. The larger the number of these entities, often called 'neurons' by analogy to the brain, the more sophisticated classifications can be produced by the classifier.**** Classifiers thus work by applying more or less sophisticated combinations of 'same' and 'different'. The

* Thinking is already sculpture.

** A good introduction into the study of classifiers is James A. Anderson, *An Introduction to Neural Networks*, MIT Press, 1995.

*** The upright posture in turn allowed humans to take breath independently of their steps, a feat that many animals cannot perform. This is turn allowed humans to develop speech. Speech in turn developed the brain.

**** A good introduction to neural nets is J. Hertz, A. Krogh & R. Palmer, *Introduction to the Theory of Neural Computation*, Addison Wesley, 1991.

distinction by a child of red and blue objects is such a classification; the distinction of compact and non-compact gauge symmetry groups in quantum theory is a more elaborate classification, but relies on the same fundamental ability.

In all classifiers, the smallest classifying units interact with each other. Often these interactions are channelled via connections, and the set is then called a *network*. In these connections, signals are exchanged, via moving objects, such as electrons or photons. Thus we arrive at the conclusion that the ability of the brain to classify the physical world, for example to distinguish moving objects interacting with each other, is a consequence of the fact that it itself consists of moving objects interacting with each other. Without a powerful classifier, humans would not have become such a successful animal species. And only the motion inside our brain allows us to talk about motion in general.

Numerous researchers are identifying the parts of the brain used when different intellectual tasks are performed. The experiments become possible using magnetic resonance imaging and other methods. Other researchers are studying how thought processes can be modelled from the brain structure. Neurology is still making regular progress. In particular, it is steadily destroying the belief that thinking is *more* than a physical process. This belief results from personal fears, as you might want to test by introspection. It will disappear as time goes by. How would you argue that thought is just a physical process?

Page 836

Challenge 1064 n

What is information?

> These thoughts did not come in any verbal formulation. I rarely think in words at all. A thought comes, and I may try to express it in words afterward.
>
> Albert Einstein

We started by stating that studying physics means to talk about motion. To talk is to transmit information. Can information be measured? Can we measure the progress of physics in this way? Is the universe made of information?

Information is the result of classification. A classification is the answer to one or to several yes–no questions. Such yes–no questions are the simplest classifications possible; they provide the basic *units* of classification, from which all others can be built. The simplest way to measure information is therefore to count the implied yes–no questions, the *bits*, leading to it. Are you able to say how many bits are necessary to define the place where you live? Obviously, the number of bits depends on the set of questions with which we start; that could be the names of all streets in a city, the set of all coordinates on the surface of the Earth, the names of all galaxies in the universe, the set of all letter combinations in the address. What is the most efficient method you can think of? A variation of the combination method is used in computers. For example, the story of this walk required about a thousand million bits. But since the amount of information in a normal letter depends on the set of questions with which we start, it is impossible to define a precise measure for information in this way.

Challenge 1065 n

The only way to measure information precisely is to take the largest possible set of questions that can be asked about a system, and to compare it with what is known about the system. In this case, the amount of unknown information is called entropy, a concept that we have already encountered. With this approach you should able to deduce yourself whether it is really possible to measure the advance of physics.

Page 217

Challenge 1066 n

Since categorization is an activity of the brain and other, similar classifiers, information as defined here is a concept that applies to the result of activities by people and by other classifiers. In short, information is produced when talking about the universe – the universe itself is *not the same* as information. There is a growing number of publications based on the opposite of this view; however, this is a conceptual short circuit. Any transmission of information implies an interaction; physically speaking, this means that any information needs *energy* for transmission and *matter* for storage. Without either of these, there is no information. In other words, the universe, with its matter and energy, has to exist *before* transmission of information is possible. Saying that the universe is made of information is as meaningful as saying that it is made of toothpaste.

The aim of physics is to give a *complete* classification of all types and examples of motion, in other words, to know everything about motion. Is this possible? Or are you able to find an argument against this endeavour? Challenge 1067 n

What is memory?

> The brain is my second favorite organ.
>
> Woody Allen

Memory is the collection of records of perceptions. The production of such records is the essential aspect of observation. Records can be stored in human memory, i.e. in the brain, or in machine memory, as in computers, or in object memory, such as notes on paper. Without memory, there is no science, no life – since life is based on the records inside the DNA – and especially, no fun, as proven by the sad life of those who lose their memory. Ref. 585

Obviously every record is an object. But under which conditions does an object qualify as a record? A signature can be the record of the agreement on a commercial transaction. A single small dot of ink is not a record, because it could have appeared by mistake, for example by an accidental blot. In contrast, it is improbable that ink should fall on paper exactly in the shape of a signature. (The simple signatures of physicians are obviously exceptions.) Simply speaking, a *record* is any object, which, in order to be copied, has to be forged. More precisely, a record is an object or a situation that cannot arise nor disappear by mistake or by chance. Our personal memories, be they images or voices, have the same property; we can usually trust them, because they are so detailed that they cannot have arisen by chance or by uncontrolled processes in our brain.

Can we estimate the probability for a record to appear or disappear by chance? Yes, we can. Every record is made of a characteristic number N of small entities, for example the number of the possible ink dots on paper, the number of iron crystals in a cassette tape, the electrons in a bit of computer memory, the silver iodide grains in a photographic negative, etc. The chance disturbances in any memory are due to internal fluctuations, also called *noise*. Noise makes the record unreadable; it can be dirt on a signature, thermal magnetization changes in iron crystals, electromagnetic noise inside a solid state memory, etc. Noise is found in all classifiers, since it is inherent in all interactions and thus in all information processing.

It is a general property that internal fluctuations due to noise decrease when the size, i.e. the number of components of the record is increased. In fact, the probability p_{mis} for

Challenge 1068 ny a misreading or miswriting of a record changes as

$$p_{\text{mis}} \sim 1/N \, , \qquad (459)$$

where N is the number of particles or subsystems used for storing it. This relation appears because, for large numbers, the so-called *normal distribution* is a good approximation of almost any process; the width of the normal distribution, which determines the probability of record errors, grows less rapidly than its integral when the number of entities is increased. (Are you able to confirm this?) Challenge 1069 ny

We conclude that any good record must be made from a *large* number of entities. The larger the number, the less sensitive the memory is to fluctuations. Now, a system of large size with small fluctuations is called a *(physical) bath*. Only baths make memories possible. In other words, every record contains a bath. We conclude that any *observation* of a system is the interaction of that system with a bath. This connection will be used several times in the following, in particular in quantum theory. When a record is produced by a machine, the 'observation' is usually called a *(generalized) measurement*. Are you able to specify the bath in the case of a person looking at a landscape? Challenge 1070 n

From the preceding discussion we can deduce a powerful conclusion: since we have such a good memory at our disposition, we can deduce that we are made of many small parts. And since records exist, the world must also be made of a large number of small parts. No microscope of any kind is needed to confirm the existence of molecules or similar small entities; such a tool is only needed to determine the *sizes* of these particles. Their existence can be deduced simply from the observation that we have memory. (Of course, another argument proving that matter is made of small parts is the ubiquity of noise.) Page 205

A second conclusion was popularized in the late 1920s by Leo Szilard. Writing a memory does not produce entropy; it is possible to store information into a memory without increasing entropy. However, entropy is produced in every case that the memory is *erased*. It turns out that the (minimum) entropy created by erasing one bit is given by Challenge 1071 ny

$$S_{\text{per erased bit}} = k \ln 2 \, , \qquad (460)$$

and the number $\ln 2 \approx 0.69$ is the natural logarithm of 2. As is well known, energy is needed to reduce the entropy of a system. Thus any system that erases memory requires energy. For example, a logical AND gate effectively erases one bit per operation. Logical thinking thus requires energy. It is also known that *dreaming* is connected with the erasing and reorganization of information. Could that be the reason that, when we are very tired, without any energy left, we do not dream as much as usual? Challenge 1072 n

Entropy is thus necessarily created when we forget. This is evident when we remind ourselves that forgetting is similar to the deterioration of an ancient manuscript. Entropy increases when the manuscript is not readable any more, since the process is irreversible and dissipative.* Another way to see this is to recognize that to clear a memory, e.g. a Ref. 588

* As Wojciech Zurek clearly explains, the entropy created inside the memory is the main reason that even Maxwell's demon cannot reduce the entropy of two volumes of gases by opening a door between them in such a way that fast molecules accumulate on one side and slow molecules accumulate on the other. (Maxwell Ref. 589

magnetic tape, we have to put energy into it, and thus increase its entropy. Conversely, *writing* into a memory can often *reduce* entropy; we remember that signals, the entities that write memories, carry negative entropy. For example, the writing of magnetic tapes usually reduces their entropy.

The capacity of the brain

> Computers are boring. They can give only answers.
> (Wrongly) attributed to Pablo Picasso

The human brain is built in such a way that its fluctuations cannot destroy its contents. The brain is well protected by the skull for exactly this reason. In addition, the brain literally grows connections, called *synapses*, between its various *neurons*, which are the cells doing the signal processing. The neuron is the basic processing element of the brain, performing the basic classification. It can only do two things: to fire and not to fire. (It is possible that the time at which a neuron fires also carries information; this question is Ref. 591, Ref. 592 not yet settled.) The neuron fires depending on its input, which comes via the synapses from hundreds of other neurons. A neuron is thus an element that can distinguish the inputs it receives into two cases: those leading to firing and those that do not. Neurons are thus classifiers of the simplest type, able only to distinguish between two situations.

Every time we store something in our long term memory, such as a phone number, new synapses are grown or the connection strength of existing synapses is changed. The connections between the neurons are much stronger than the fluctuations in the brain. Only strong disturbances, such as a blocked blood vessel or a brain lesion, can destroy neurons and lead to loss of memory.

As a whole, the brain provides an extremely efficient memory. Despite intense efforts, engineers have not yet been able to build a memory with the capacity of the brain in the same volume. Let us estimated this memory capacity. By multiplying the number of neurons, about 10^{11}, by the average number of synapses per neuron, about 100, and also by the estimated number of bits stored in every synapse, about 10, we arrive at a storage capacity for the brain of about

$$M_{\text{rewritable}} \approx 10^{14}\,\text{bit} \approx 10^{4}\,\text{GB}\,. \qquad (461)$$

(One *byte*, abbreviated B, is the usual name for eight bits of information.) Note that evolution has managed to put as many neurons in the brain as there are stars in the galaxy, and that if we add all the synapse lengths, we get a total length of about 10^{11} m, which corresponds to the distance to from the Earth to the Sun. Our brain truly is *astronomically* complex.

In practice, the capacity of the brain seems almost without limit, since the brain frees memory every time it needs some new space, by *forgetting* older data, e.g. during sleep. Note that this standard estimate of 10^{14} bits is not really correct! It assumes that the only

---

had introduced the 'demon' in 1871, to clarify the limits posed by nature to the gods.) This is just another way to rephrase the old result of Leo Szilard, who showed that the measurements by the demon create more Ref. 590 entropy than they can save. And every measurement apparatus contains a memory.

To play being Maxwell's demon, click on the http://www.wolfenet.com/~zeppelin/maxwell.htm website.

component storing information in the brain is the synapse strength. Therefore it only measures the *erasable* storage capacity of the brain. In fact, information is also stored in the structure of the brain, i.e. in the exact configuration in which every cell is connected to other cells. Most of this structure is fixed at the age of about two years, but it continues to develop at a lower level for the rest of human life. Assuming that for each of the N cells with n connections there are $f\,n$ connection possibilities, this *write once* capacity of the brain can be estimated as roughly $N\sqrt{f n}\, f n \log f n$ bits. For $N = 10^{11}$, $n = 10^2$, $f = 6$, this gives

Challenge 1073 ny

$$M_{\text{writeonce}} \approx 10^{16}\,\text{bit} \approx 10^6\,\text{GB}\ . \tag{462}$$

Recent measurements confirmed that bilingual persons, especially early bilinguals, have a higher density of grey mass in the small parietal cortex on the left hemisphere of the brain. This is a region mainly concerned with language processing. The brain thus makes also use of structural changes for optimized storage and processing.

Incidentally, even though the brains of sperm whales and of elephants can be five to six times as heavy as those of humans, the number of neurons and connections, and thus the capacity, is lower than for humans.

Sometimes it is claimed that people use only between 5% or 10% of their brain capacity. This myth, which goes back to the nineteenth century, would imply that it is possible to measure the actual data stored in the brain and compare it with its capacity to an impossible accuracy. Alternatively, the myth implies that the processing capacity can be measured. It also implies that nature would develop and maintain an organ with 90 % overcapacity, wasting all the energy and material to build, repair and maintain it. The myth is wrong.

The large storage capacity of the brain also shows that human memory is filled by the environment and is not inborn: one human ovule plus one sperm have a mass of about 1 mg, which corresponds to about $3{\cdot}10^{16}$ atoms. Obviously, fluctuations make it impossible to store 10^{16} bits in it. In fact, nature stores only about $3 \cdot 10^9$ bits in the genes of an ovule, using 10^7 atoms per bit. In contrast, a typical brain has a mass of 1.5 to 2 kg, containing about 5 to $7 \cdot 10^{25}$ atoms, which makes it as efficient as the ovule. The difference between the number of bits in human DNA and those in the brain nicely shows that almost all information stored in the brain is taken from the environment; it cannot be of genetic origin, even allowing for smart decompression of stored information.

In total, all these tricks used by nature result in the most powerful classifier yet known. Are there any limits to the brain's capacity to memorize and to classify? With the tools that humans have developed to expand the possibilities of the brain, such as paper, writing and printing to help memory, and the numerous tools available to simplify and to abbreviate classifications explored by mathematicians, brain classification is only limited by the time spent practising it.Without tools, there are strict limits, of course. The two-millimetre thick cerebral cortex of humans has a surface of about four sheets of A4 paper, a chimpanzee's yields one sheet and a monkey's is the size of a postcard. It is estimated that the total intellectually accessible memory is of the order of

Ref. 593

$$M_{\text{intellectual}} \approx 1\,\text{GB}\ , \tag{463}$$

though with a large experimental error.

The brain is also unparalleled in its processing capacity. This is most clearly demonstrated by the most important consequence deriving from memory and classification: thought and language. Indeed, the many types of thinking or language we use, such as comparing, distinguishing, remembering, recognizing, connecting, describing, deducing, explaining, imagining, etc., all describe different ways to classify memories or perceptions. In the end all thinking and talking directly or indirectly classify observations. But how far are computers from achieving this! To talk to a computer program, such as to the famous program Eliza and its successors that mimic a psychoanalyst, is stilla disappointing experience. To understand the reasons for this slow development, we ask:

What is language?

> Reserve your right to think, for even to think wrongly is better than not to think at all.
>
> Hypatia of Alexandria (*c.* 355–415)

> Ein Satz kann nur sagen, *wie* ein Ding ist, nicht *was* es ist.*
>
> Ludwig Wittgenstein, *Tractatus*, 3.221

Language possibly is the most wonderful gift of human nature. Using the ability to produce sounds and to put ink on paper, people attach certain *symbols*,** also called *words* or *terms* in this context, to the many partitions they specify with the help of their thinking. Such a categorization is then said to define a *concept* or *notion*, and is set in *italic typeface* in this text. A standard set of concepts forms a language.*** In other words, a *(human) language* is a standard way of symbolic interaction between people.**** There are human languages based on facial expressions, on gestures, on spoken words, on whistles, on written words, and more. The use of *spoken* language is considerably younger than the human species; it seems that it appeared only about two hundred thousand years ago. Written language is even younger, namely only about six thousand years old. But the set of concepts used, the *vocabulary*, is still expanding. For humans, the understanding of language begins soon after birth (perhaps even before), the active use begins at around a

* Propositions can only say *how* things are, not *what* they are.

** A symbol is a type of *sign*, i.e. an entity associated by some convention to the object it refers. Following Charles Peirce (1839–1914) – see http://www.peirce.org – the most original philosopher born in the United States, a symbol differs from an *icon* (or *image*) and from an *index*, which are also attached to objects by convention, in that it does not resemble the object, as does an icon, and in that it has no contact with the object, as is the case for an index.

*** The recognition that language is based on a partition of ideas, using the various differences between them to distinguish them from each other, goes back to the Swiss thinker Ferdinand de Saussure (1857–1913), who is regarded as the founder of linguistics. His textbook *Cours de linguistique générale*, Editions Payot, Paris, 1985, has been the reference work of the field for over half a century. Note that Saussure, in contrast to Peirce, prefers the term 'sign' to 'symbol', and that his definition of the term 'sign' includes also the object to which it refers.

**** For slightly different definitions and a wealth of other interesting information about language, see the beautiful book by David Crystal, *The Cambridge Encyclopedia of Language*, Cambridge University Press, 1987.

year of age, the ability to read can start as early as two, and personal vocabulary continues to grow as long as curiosity is alive.

Physics being a lazy way to chat about motion, it needs language as an essential tool. Of the many aspects of language, from literature to poetry, from jokes to military orders, from expressions of encouragement, dreams, love and emotions, physics uses only a small and rather special segment. This segment is defined by the inherent restriction to talk about motion. Since motion is an observation, i.e. an interaction with the environment that several people experience in the same way, this choice puts a number of restrictions on the contents – the vocabulary – and on the form – the grammar – of such discussions.

For example, from the definition that observations are shared by others, we get the requirement that the statements describing them must be translatable into all languages. But when can a statement be translated? On this question two extreme points of view are possible: the first maintains that *all* statements can be translated, since it follows from the properties of human languages that each of them can express every possible statement. In this view, only sign systems that allow one to express the complete spectrum of human messages form a *human language*. This property distinguishes spoken and sign language from animal languages, such as the signs used by apes, birds or honey bees, and also from computer languages, such as Pascal or C. With this meaning of language, all statements can be translated by definition.

It is more challenging for a discussion to follow the opposing view, namely that precise translation is possible only for those statements which use terms, word types and grammatical structures found in *all* languages. Linguistic research has invested considerable effort in the distillation of phonological, grammatical and semantic *universals*, as they are called, from the 7000 or so languages thought to exist today.*

The investigations into the *phonological* aspect, which showed for example that every language has at least two consonants and two vowels, does not provide any material for the discussion of translation.** Studying the *grammatical* (or *syntactic*) aspect, one finds that all languages use smallest elements, called 'words', which they group into sentences. They all have pronouns for the first and second person, 'I' and 'you', and always contain nouns and verbs. All languages use *subjects* and *predicates* or, as one usually says, the three entities *subject*, *verb* and *object*, though not always in this order. Just check the languages

Challenge 1074 e
you know.

On the *semantic* aspect, the long list of lexical universals, i.e. words that appear in all languages, such as 'mother' or 'Sun', has recently been given a structure. The linguist Anna Wierzbicka performed a search for the building blocks from which all concepts can be built. She looked for the definition of every concept with the help of simpler ones, and continued doing so until a fundamental level was reached that cannot be further reduced.

* A comprehensive list with 6 800 languages (and with 41 000 language and dialect names) can be found on the world wide website by Barbara Grimes, *Ethnologue – Languages of the World*, to be found at the address http://www.ethnologue.com or in the printed book of the same name.

It is estimated that 15 000 ± 5 000 languages have existed in the past.

Nevertheless, in today's world, and surely in the sciences, it is often sufficient to know one's own language plus English. Since English is the language with the largest number of words, learning it well is a greater

Ref. 594
challenge than learning most other languages.

** Studies explore topics such as the observation that in many languages the word for 'little' contains an 'i' (or high pitched 'e') sound: petit, piccolo, klein, tiny, pequeño, chiisai; exceptions are: small, parvus.

Table 50 The semantic primitives, following Anna Wierzbicka

I, you, someone, something, people	[substantives]
this, the same, one, two, all, much/many	[determiners and quantifiers]
know, want, think, feel, say	[mental predicates]
do, happen	[agent, patient]
good, bad	[evaluative]
big, small	[descriptors]
very	[intensifier]
can, if (would)	[modality, irrealis]
because	[causation]
no (not)	[negation]
when, where, after (before), under (above)	[time and place]
kind of, part of	[taxonomy, partonomy]
like	[hedge/prototype]

The set of concepts that are left over are the primitives. By repeating this exercise in many languages, Wierzbicka found that the list is the same in all cases. She thus had discovered *universal semantic primitives*. In November 1992, the list contained the terms given in Table 50.

Following the life-long research of Anna Wierzbicka and her research school, all these concepts exist in all languages of the world studied so far.* They have defined the meaning of each primitive in detail, performed consistency checks and eliminated alternative approaches. They have checked this list in languages from all language groups, in languages from all continents, thus showing that the result is valid everywhere. In every language all other concepts can be defined with the help of the semantic primitives. Ref. 595

Simply stated, learning to speak means learning these basic terms, learning how to combine them and learning the names of these composites. The definition of language given above, namely as a means of communication that allows one to express everything one wants to say, can thus be refined: a *human language* is any set of concepts that includes the universal semantic primitives.

For physicists – who aim to talk in as few words as possible – the list of semantic primitives has three facets. First, the approach is appealing, as it is similar to physics' own aim: the idea of primitives gives a structured summary of everything that can be said, just as the atomic elements structure all objects that can be observed. Second, the list of primitives can be structured. In fact, the list of primitives can be divided into two groups: one group contains all terms describing motion (do, happen, when, where, feel, small, etc. – probably a term from the semantic field around light or colour should be added) and the other group contains all terms necessary to talk about abstract sets and

* It is easy to imagine that this research steps on the toes of many people. A list that maintains that 'true', 'good', 'creation', 'life', 'mother' or 'god' are composite will elicit violent reactions, despite the correctness of the statements. Indeed, some of these terms were added in the 1996 list, which is somewhat longer. In addition, a list that maintains that we only have about thirty basic concepts in our heads is taken by many to be offensive.

relations (this, all, kind of, no, if, etc.). Even for linguistics, aspects of motion and logical concepts are the basic entities of human experience and human thinking. To bring the issue to a point, the semantic primitives contain the basic elements of physics and the basic elements of mathematics. All humans are thus both physicists and mathematicians. The third point is that the list of primitives is too long. The division of the list into two groups directly suggests shorter lists; we just have to ask physicists and mathematicians for concise summaries of their respective fields. To appreciate this aim, try to define what
Challenge 1075 d 'if' means, or what an 'opposite' is – and explore your own ways of reducing the list.

Reducing the list of primitives is also one of our aims in this adventure. We will explore the mathematical group of primitives in this intermezzo; the physical group will occupy us in the rest of our adventure. However, a shorter list of primitives is not sufficient. Our goal is to arrive at a list consisting of only *one* basic concept. Reaching this goal is not simple, though. First, we need to check whether the set of classical physical concepts that we have discovered so far is complete. Can classical physical concepts describe *all* observations? The second part of our adventure is devoted to this question. The second task is to reduce the list. This task is not straightforward; we have already discovered that physics is based on a circular definition: in Galilean physics, space and time are defined
Page 144 using matter, and matter is defined using space and time. We will need quite some effort to overcome this obstacle. The third part of this text tells the precise story. After numerous adventures we will indeed discover a basic concept on which all other concepts are based.

We can summarize all the above-mentioned results of linguistics in the following way. By constructing a statement made only of subject, verb and object, consisting only of nouns and verbs, using only concepts built from the semantic primitives, we are sure that it can be translated into all languages. This explains why physics textbooks are often so boring: the authors are often too afraid to depart from this basic scheme. On the other hand, research has shown that such straightforward statements are not restrictive: with them one can say everything that can be said.

> Every word was once a poem.
>
> Ralph Waldo Emerson*

What is a concept?

> Concepts are merely the results, rendered permanent by language, of a previous process of comparison.
>
> William Hamilton

There is a group of people that has taken the strict view on translation and on precision to the extreme. They build all concepts from an even smaller set of primitives, namely only two: 'set' and 'relation', and explore the various possible combinations of these two concepts, studying their classifications. Step by step, this radical group, commonly called *mathematicians*, came to define with full precision concepts such as numbers, points, curves, equations, symmetry groups and more. The construction of these concepts is summarized partly in the following and partly in Appendix D.

* Ralph Waldo Emerson (1803–1882), US-American essayist and philosopher.

However, despite their precision, in fact precisely because of it, no mathematical concept talks about nature or about observations.* Therefore the study of motion needs other, more useful concepts. What properties must a useful concept have? For example, what is 'freedom' or what is a 'parachute'? Obviously, a useful concept implies a list of its parts, its aspects and their internal relations, as well as their relation to the exterior world. Thinkers in various fields, from philosophy to politics, agree that the definition of any *concept* requires:

- explicit and fixed content,
- explicit and fixed limits,
- explicit and fixed domain of application.

The inability to state these properties or keep them fixed is often the easiest way to distinguish *crackpots* from more reliable thinkers. Unclearly defined terms, which thus do not qualify as concepts, regularly appear in myths, e.g. 'dragon' or 'sphinx', or in ideologies, e.g. 'worker' or 'soul'. Even physics is not immune. For example, we will discover later that neither 'universe' nor 'creation' are concepts. Are you able to argue the case? Challenge 1076 n

But the three defining properties of any concepts are interesting in their own right. Explicit content means that concepts are built one onto another. In particular, the most fundamental concepts appear to be those that have no parts and no external relations, but only internal ones. Can you think of one? Only the last part of this walk will uncover the final word on the topic. Challenge 1077 n

The requirements of explicit limits and explicit contents also imply that all concepts describing nature are *sets*, since sets obey the same requirements. In addition, explicit domains of application imply that all concepts also are *relations*.* Since mathematics is based on the concepts of 'set' and of 'relation', one follows directly that mathematics can provide the *form* for any concept, especially whenever high precision is required, as in the study of motion. Obviously, the *content* of the description is only provided by the study of nature itself; only then do concepts become useful.

In the case of physics, the search for sufficiently precise concepts can be seen as the single theme structuring the long history of the field. Regularly, new concepts have been proposed, explored in all their properties, and tested. Finally, concepts are rejected or adopted, in the same way that children reject or adopt a new toy. Children do this unconsciously, scientists do it consciously, using language.** For this reason, concepts are

* Insofar as one can say that mathematics is based on the concepts of 'set' and 'relation', which are based on experience, one can say that mathematics explores a section of reality, and that its concepts are *derived* from experience. This and similar views of mathematics are called *platonism*. More concretely, platonism is the view that the concepts of mathematics exist *independently* of people, and that they are discovered, and not created, by mathematicians.

In short, since mathematics makes use of the brain, which is a physical system, actually *mathematics is applied physics*.

* We see that every physical concept is an example of a (mathematical) *category*, i.e. a combination of objects and mappings. For more details about categories, with a precise definition of the term, see page 603.

** Concepts formed unconsciously in our early youth are the most difficult to define precisely, i.e. with language. Some who were unable to define them, such as the Prussian philosopher Immanuel Kant (1724–1804) used to call them 'a priori' concepts (such as 'space' and 'time') to contrast them with the more clearly defined 'a posteriori' concepts. Today, this distinction has been shown to be unfounded both by the study of child psychology (see the footnote on page 587) and by physics itself, so that these qualifiers are therefore not used in our walk.

universally intelligible.

Note that the concept '*concept*' itself is not definable independently of experience; a concept is something that helps us to act and react to the world in which we live. Moreover, concepts do not live in a world separate from the physical one: every concept requires memory from its user, since the user has to remember the way in which it was formed; therefore every concept needs a material support for its use and application. Thus all thinking and all science is fundamentally based on experience.

In conclusion, all concepts are based on the idea that nature is made of related parts. This idea leads to complementing couples such as 'noun–verb' in linguistics, 'set–relation' or 'definition–theorem' in mathematics, and 'aspect of nature–pattern of nature' in physics. These couples constantly guide human thinking, from childhood onwards, as developmental psychology can testify.

What are sets? What are relations?

> Alles, was wir sehen, könnte auch anders sein. Alles, was wir überhaupt beschreiben können, könnte auch anders sein. Es gibt keine Ordnung der Dinge a priori.*
>
> Ludwig Wittgenstein, *Tractatus*, 5.634

Defining sets and defining relations are the two fundamental acts of our thinking. This can be seen most clearly in any book about mathematics; such a book is usually divided into paragraphs labelled 'definition', 'theorem', 'lemma' and 'corollary'. The first type of paragraph defines concepts, i.e. defines sets, and the other three types of paragraphs express relations, i.e. connections between these sets. *Mathematics* is thus the exploration of the possible symbolic concepts and their relations. Mathematics is the science of symbolic necessities.

Sets and relations are tools of classification; that is why they are also the tools of any bureaucrat. (See Figure 270.) This class of humans is characterized by heavy use of paper clips, files, metal closets, archives – which all define various types of sets – and by the extensive use of numbers, such as reference numbers, customer numbers, passport numbers, account numbers, law article numbers – which define various types of relations between the items, i.e. between the elements of the sets.

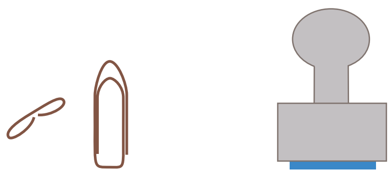


Figure 270 Devices for the definition of sets (left) and of relations (right)

Both the concepts of set and of relation express, in different ways, the fact that nature can be *described*, i.e. that it can be classified into parts that form a whole. The act of grouping together aspects of experience, i.e. the act of classifying them, is expressed in formal language by saying that a set is defined. In other words, a *set* is a collection of *elements* of our thinking. Every set distinguishes the elements from each other and from the set itself. This definition of 'set' is called the *naive* definition. For physics, the

* Whatever we see could be other than it is. Whatever we can describe at all could be other than it is. There is no a priori order of things.

Table 51 The defining properties of a set – the ZFC axioms

The axioms of Zermelo–Fraenkel–C set theory
▪ Two sets are equal if and only if they have the same elements. (Axiom of extensionality)
▪ The empty set is a set. (Axiom of the null set)
▪ If x and y are sets, then the unordered pair $\{x, y\}$ is a set. (Axiom of unordered pairs)
▪ If x is a set of sets, the union of all its members is a set. (Union or sum set axiom)
▪ The entity $\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \{\{\{\varnothing\}\}\}, ...\}$ is a set – in other words, infinite collections such as the natural numbers are sets. (Axiom of infinity)
▪ An entity defined by all elements having a given property is a set, provided this property is reasonable; some important technicalities defining 'reasonable' are necessary. (Axiom of replacement)
▪ The entity y of all subsets of x is also a set, called the power set. (Axiom of the power set)
▪ A set is not an element of itself – plus some technicalities. (Axiom of regularity)
▪ Picking elements from a list of sets allows one to construct a new set – plus technicalities. (Axiom of choice)

definition is sufficient, but you won't find many who will admit this. In fact, mathematicians have refined the definition of the concept 'set' several times, because the naive definition does not work well for infinite sets. A famous example is the story about sets which do not contain themselves. Obviously, any set is of two sorts: either it contains itself or it does not. If we take the set of all sets that do *not* contain themselves, to which sort does it belong? Challenge 1078 n

To avoid problems with the concept of 'set', mathematics requires a precise definition. The first such definition was given by the German mathematician Ernst Zermelo (b. 1871 Berlin, d. 1951 Freiburg i.B.) and the German–Israeli mathematician Adolf/Abraham Fraenkel (b. 1891 München, d. 1965 Jerusalem). Later, the so-called *axiom of choice* was added, in order to make it possible to manipulate a wider class of infinite sets. The result of these efforts is called the ZFC definition.* From this basic definition we can construct all mathematical concepts used in physics. From a practical point of view, it is sufficient to keep in mind that for the whole of physics, the naive definition of a set is equivalent to the precise ZFC definition, actually even to the simper ZF definition. Subtleties appear only for some special types of infinite sets, but these are not used in physics. In short, from the basic, naive set definition we can construct all concepts used in physics.

Ref. 597 The naive set definition is far from boring. To satisfy two people when dividing a cake, we follow the rule: I cut, you choose. The method has two properties: it is *just*, as everybody thinks that they have the share that they deserve, and it is *fully satisfying*, as everybody has the feeling that they have at least as much as the other. What rule is needed for

* A global overview of axiomatic set theory is given by Paul J. Cohen & Reuben Hersch, Non-Cantorian set theory, *Scientific American* 217, pp. 104–116, 1967. Those were the times when *Scientific American* was a quality magazine.

Ref. 596 Other types of entities, more general than standard sets, obeying other properties, can also be defined, and are also subject of (comparatively little) mathematical research. To find an example, see the section on cardinals later on. Page 602 Such more general entities are called classes whenever they contain at least one set. Can you give an example? Challenge 1079 n In the third part of our mountain ascent we will meet physical concepts that are described neither by sets nor by classes, containing no set at all. That is were the real fun starts.

Challenge 1080 d three people? And for four?

Apart from defining sets, every child and every brain creates links between the different aspects of experience. For example, when it hears a voice, it automatically makes the connection that a human is present. In formal language, connections of this type are called *relations*. Relations connect and differentiate elements along other lines than sets: the two form a complementing couple. Defining a set unifies many objects and at the same time divides them into two: those belonging to the set and those that do not; defining a (binary) relation unifies elements two by two and divides them into many, namely into the many couples it defines.

Sets and relations are closely interrelated concepts. Indeed, one can define (mathematical) relations with the help of sets. A *(binary) relation* between two sets X and Y is a subset of the product set, where the *product set* or *Cartesian product* $X \times Y$ is the set of all ordered pairs (x, y) with $x \in X$ and $y \in Y$. An *ordered pair* (x, y) can easily be defined with the help of sets. Can you find out how? Challenge 1081 n For example, in the case of the relation 'is wife of', the set X is the set of all women and the set Y that of all men; the relation is given by the list all the appropriate ordered pairs, which is much smaller than the product set, i.e. the set of all possible woman–man combinations.

It should be noted that the definition of relation just given is not really complete, since every construction of the concept 'set' already contains certain relations, such as the relation 'is element of.' It does not seem to be possible to reduce either one of the concepts 'set' or 'relation' completely to the other one. This situation is reflected in the physical cases of sets and relations, such as space (as a set of points) and distance, which also seem impossible to separate completely from each other. In other words, even though mathematics does not pertain to nature, its two basic concepts, sets and relations, are taken from nature. In addition, the two concepts, like those of space-time and particles, are each defined with the other.

Infinity

Mathematicians soon discovered that the concept of 'set' is only useful if one can also call collections such as $\{0, 1, 2, 3...\}$, i.e. of the number 0 and all its successors, a 'set'. To achieve this, one property in the Zermelo–Fraenkel list defining the term 'set' explicitly specifies that this collection can be called a set. (In fact, also the axiom of replacement states that sets may be infinite.) Infinity is thus put into mathematics and into the tools of our thought right at the very beginning, in the definition of the term 'set'. When describing nature, with or without mathematics, we should never forget this fact. A few additional points about infinity should be of general knowledge to any expert on motion.

Only *sets* can be infinite. And sets have parts, namely their elements. When a thing or a concept is called 'infinite' one can *always* ask and specify what its parts are: for space the parts are the points, for time the instants, for the set of integers the integers, etc. An indivisible or a finitely divisible entity cannot be called infinite.*

A set is infinite if there is a function from it into itself that is *injective* (i.e. different elements map to different results) but not *onto* (i.e. some elements do not appear as images

* Therefore, most gods, being concepts and thus sets, are either finite or, in the case where they are infinite, they are divisible. It seems that only polytheistic world views are not disturbed by this conclusion.

of the map); e.g. the map $n \mapsto 2n$ shows that the set of integers is infinite. Infinity also can be checked in another way: a set is infinite if it remains so also after removing one element, even repeatedly. We just need to remember that the empty set is *finite*.

There are *many types* of infinities, all of different sizes.* This important result was discovered by the Danish-Russian-German mathematician Georg Cantor (1845–1918). He showed that from the countable set of natural numbers one can construct other infinite sets which are not countable. He did this by showing that the *power set* $P(\omega)$, namely the set of all subsets, of a countably infinite set is infinite, but *not* countably infinite. Sloppily speaking, the power set is 'more infinite' than the original set. The real numbers **R**, to be defined shortly, are an example of an uncountably infinite set; there are many more of
Challenge 1082 n them than there are natural numbers. (Can you show this?) However, *any* type of infinite set contains at least one subset which is countably infinite.

Even for an infinite set one can define size as the number of its elements. Cantor called this the *cardinality* of a set. The cardinality of a finite set is simply given by the number of its elements. The cardinality of a power set is 2 exponentiated by the cardinality of the set. The cardinality of the set of integers is called $\aleph_0$, pronounced 'aleph zero', after the
Page 1054 first letter of the Hebrew alphabet. The smallest *uncountable* cardinal is called $\aleph_1$. The next cardinal is called $\aleph_2$ etc. A whole branch of mathematics is concerned with the manipulation of these infinite 'numbers'; addition, multiplication, exponentiation are easily defined. For some of them, even logarithms and other functions make sense.**

The cardinals defined in this way, including $\aleph_n$, $\aleph_\omega$, $\aleph_{\aleph_\aleph}$ are called accessible, because since Cantor, people have defined even larger types of infinities, called *inaccessible*. These numbers (inaccessible cardinals, measurable cardinals, supercompact cardinals, etc.) need additional set axioms, extending the zfc system. Like the ordinals and the cardinals, they form examples of what are called *transfinite* numbers.

The real numbers have the cardinality of the power set of the integers, namely $2^{\aleph_0}$.
Challenge 1083 n Can you show this? The result leads to the famous question: Is $\aleph_1 = 2^{\aleph_0}$ or not? The statement that this be so is called the *continuum hypothesis* and was unproven for several generations. The surprising answer came in 1963: the usual definition of the concept of
Ref. 598 set is not specific enough to fix the answer. By specifying the concept of set in more detail, with additional axioms – remember that axioms are defining properties – you can make the continuum hypothesis come out either right or wrong, as you prefer.

Another result of research into transfinites is important: for every definition of a type of infinite cardinal, it seems to be possible to find a larger one. In everyday life, the idea of infinity is often used to stop discussions about size: 'My big brother is stronger than yours.' 'But mine is infinitely stronger than yours!' Mathematics has shown that questions on size do continue afterwards: 'The strength of my brother is the power set of that of
Ref. 599 yours!' Rucker reports that mathematicians conjecture that there is no possible nor any conceivable end to these discussions.

For physicists, a simple question appears directly. Do infinite quantities exist in nature? Or better, is it necessary to use infinite quantities to describe nature? You might want to
Challenge 1084 e clarify your own opinion on the issue. It will be settled during the rest of our adventure.

* In fact, there such a huge number of types of infinities that none of these infinities itself actually describes this number. Technically speaking, there are as many infinities as there are ordinals.
** Many results are summarized in the excellent and delightful paperback by Rudy Rucker, *Infinity and*

Functions and structures

Which relations are useful to describe patterns in nature? A typical example is 'larger stones are heavier'. Such a relation is of a specific type: it relates one specific value of an observable 'volume' to one specific value of the observable 'weight'. Such a one-to-one relation is called a *(mathematical) function* or *mapping*. Functions are the most specific types of relations; thus they convey a maximum of information. In the same way as numbers are used for observables, functions allow easy and precise communication of relations between observations. All physical rules and 'laws' are therefore expressed with the help of functions and, since physical 'laws' are about measurements, functions of numbers are their main building blocks.

A *function* f, or *mapping*, is a thus binary relation, i.e. a set $f = \{(x, y)\}$ of ordered pairs, where for every value of the first element x, called the *argument*, there is only *one* pair (x, y). The second element y is called the *value* of the function at the argument x. The set X of all arguments x is called the *domain of definition* and the set Y of all second arguments y is called the *range* of the function. Instead of $f = \{(x, y)\}$ one writes

$$f : X \to Y \quad \text{and} \quad f : x \mapsto y \quad \text{or} \quad y = f(x) \;, \tag{464}$$

where the type of arrow – with initial bar or not – shows whether we are speaking about sets or about elements.

We note that it is also possible to use the couple 'set' and 'mapping' to define all mathematical concepts; in this case a relation is defined with the help of mappings. A modern school of mathematical thought formalized this approach by the use of (mathematical) *categories*, a concept that includes both sets and mappings on an equal footing in its definition.*

To think and talk more clearly about nature, we need to define more specialized concepts than sets, relations and functions, because these basic terms are too general. The most important concepts derived from them are operations, algebraic structures and numbers.

A *(binary) operation* is a function that maps the Cartesian product of two copies of a set X into itself. In other words, an operation w takes an ordered couple of arguments $x \in X$ and assigns to it a value $y \in X$:

$$w : X \times X \to X \quad \text{and} \quad w : (x, x) \mapsto y \;. \tag{465}$$

Challenge 1085 n Is division of numbers an operation in the sense just defined?

the Mind – the Science and Philosophy of the Infinite, Bantam, Toronto, 1983.

* A *category* is defined as a collection of objects and a collection of 'morphisms', or mappings. Morphisms can be composed; the composition is associative and there is an identity morphism. The strange world of category theory, sometimes called the abstraction of all abstractions, is presented in F. William Lawvere & Stephen H. Schanuel, *Conceptual Mathematics: a First Introduction to Categories*, Cambridge University Press, 1997.

Note that every category contains a set; since it is unclear whether nature contains sets, as we will discuss on page 633, it is questionable whether categories will be useful in the unification of physics, despite their intense and abstract charm.

Now we are ready to define the first of three basic concepts of mathematics. An *algebraic structure*, also called an *algebraic system*, is (in the most restricted sense) a set together with certain operations. The most important algebraic structures appearing in physics are groups, vector spaces, and algebras.

Page 1107

In addition to algebraic structures, mathematics is based on *order structures* and on *topological structures*. Order structures are building blocks of numbers and necessary to define comparisons of any sort. Topological structures are built, via subsets, on the concept of neighbourhood. They are necessary to define continuity, limits, dimensionality, topological spaces and manifolds.

Page 1116

Obviously, most mathematical structures are combinations of various examples of these three basic structure types. For example, the *system* of real numbers is given by the *set* of real numbers with the *operations* of addition and multiplication, the *order relation* 'is larger than' and a *continuity* property. They are thus built by combining an algebraic structure, an order structure and a topological structure. Let us delve a bit into the details.

Ref. 600

Numbers

Which numbers are multiplied by six when their last digit is taken away and transferred to the front?

Challenge 1086 n

Numbers are the oldest mathematical concept and are found in all cultures. The notion of number, in Greek ἀριθμός, has been changed several times. Each time the aim was to include wider classes of objects, but always retaining the general idea that numbers are entities that can be added, subtracted, multiplied and divided.

The modern way to write numbers, as e.g. in $12\,345\,679 \cdot 45 = 666\,666\,666$, is essential for science.* It can be argued that the lack of a good system for writing down and for calculating with numbers delayed the progress of science by several centuries. (By the way, the same can be said for the affordable mass reproduction of written texts.)

The simplest numbers, $0, 1, 2, 3, 4, \ldots$, are usually seen as being taken directly from experience. However, they can also be constructed from the notions of 'relation' and 'set'. One of the many possible ways to do this (can you find another?) is by identifying a natural number with the set of its predecessors. With the relation 'successor of', abbreviated S, this definition can be written as

Challenge 1087 n

$$0 := \varnothing \quad , \quad 1 := S\,0 = \{0\} = \{\varnothing\} \quad ,$$
$$2 := S\,1 = \{0,1\} = \{\varnothing, \{\varnothing\}\} \quad \text{and} \quad n+1 := S\,n = \{0, \ldots, n\} \,. \qquad (466)$$

This set, together with the binary operations 'addition' and 'multiplication,' constitutes the algebraic system $N = (N, +, \cdot, 1)$ of the *natural numbers*. For all number systems the algebraic system and the set are often sloppily designated by the same symbol. The algebraic system N is a so-called semi-ring, as explained in Appendix D. (Some authors prefer not to count the number zero as a natural number.) Natural numbers are fairly useful.

Page 1098

* However, there is no need for written numbers for doing mathematics, as shown by MARCIA ASCHER, *Ethnomathematics – A Multicultural View of Mathematical Ideas*, Brooks/Cole, 1991.

Table 52 Some large numbers

Number	Example in nature
Around us	
1	number of angels that can be in one place at the same time, following Thomas Aquinas Ref. 601
8	number of times a newspaper can be folded in alternate perpendicular directions
12	largest number of times a paper strip has been folded in the same direction
20	number of digits in precision measurements that will probably never be achieved
34, 55, 89	petals of common types of daisy and sunflower Ref. 602
57	faces of a diamond with brilliant cut
2000	stars visible in the night sky
10^5	leaves of a tree (10 m beech)
6 to 7 $\cdot 10^9$	humans in the year 2000
10^{17}	ants in the world
c. 10^{20}	number of snowflakes falling on the Earth per year
c. 10^{24}	grains of sand in the Sahara desert
10^{22}	stars in the universe
10^{25}	cells on Earth
$1.1 \cdot 10^{50}$	atoms making up the Earth ($6370^3\,\mathrm{km}^3 \cdot 4 \cdot 3.14/3 \cdot 5500\,\mathrm{kg/m}^3 \cdot 30\,\mathrm{mol/kg} \cdot 6 \cdot 10^{23}\,/\mathrm{mol}$)
10^{81}	atoms in the visible universe
10^{90}	photons in the visible universe
10^{169}	number of atoms fitting in the visible universe
10^{244}	number of space-time points inside the visible universe
Information	
51	record number of languages spoken by one person
c. 5000	words spoken on an average day by a man
c. 7000	words spoken on an average day by a woman
c. 7000	number of languages on Earth
c. 350 000	words of the English language (more than any other language, with the possible exception of German)
c. 2 000 000	number of scientists on Earth around the year 2000
$3 \cdot 10^8$	words spoken during a lifetime (2/3 time awake, 30 words per minute)
$4 \cdot 10^9$	pulses exchanged between both brain halves every second
10^9	words heard and read during a lifetime
10^{17}	image pixels seen in a lifetime ($3 \cdot 10^9\,\mathrm{s} \cdot (1/15\,\mathrm{ms}) \cdot 2/3$ (awake) $\cdot 10^6$ (nerves to the brain) Ref. 603
10^{19}	bits of information processed in a lifetime (the above times 32)
c. $5 \cdot 10^{12}$	printed words available in (different) books around the world (*c.* $100 \cdot 10^6$ books consisting of 50 000 words)
$2^{10} \cdot 3^7 \cdot 8! \cdot 12!$	

Number	Example in nature
$= 4.3 \cdot 10^{19}$	possible positions of the 3 × 3 × 3 Rubik's Cube Ref. 604
$5.8 \cdot 10^{78}$	possible positions of the 4 × 4 × 4 Rubik-like cube
$5.6 \cdot 10^{117}$	possible positions of the 5 × 5 × 5 Rubik-like cube
c. 10^{200}	possible games of chess
c. 10^{800}	possible games of go
c. 10^{10^7}	possible states in a personal computer
Parts of us	
600	numbers of muscles in the human body, of which about half are in the face
$150\,000 \pm 50\,000$	hairs on a healthy head
900 000	neurons in the brain of a grasshopper
$126 \cdot 10^6$	light sensitive cells per retina (120 million rods and 6 million cones)
10^{10} to 10^{11}	neurons in the human brain
$> 10^{16}$	memory bits in the human brain
$500 \cdot 10^6$	blinks of the eye during a lifetime (about once every four seconds when awake)
$300 \cdot 10^6$	breaths taken during human life
$3 \cdot 10^9$	heart beats during a human life
$3 \cdot 10^9$	letters (base pairs) in haploid human DNA
$6.1 \cdot 10^9$	bits in a compact disk
$1 \cdot 10^{11}$	humans who have ever lived
$10^{15\pm1}$	cells in the human body
$10^{16\pm1}$	bacteria carried in the human body

The system of *integers* $Z = (..., -2, -1, 0, 1, 2, ..., +, \cdot, 0, 1)$ is the minimal ring that is an extension of the natural numbers. The system of *rational numbers* $Q = (Q, +, \cdot, 0, 1)$ is the minimal field that is an extension of the ring of the integers. (The terms 'ring' and 'field' are explained in Appendix D.) Page 1098 The system of *real numbers* $R = (R, +, \cdot, 0, 1, >)$ is the minimal extension of the rationals that is continuous and totally ordered. (For the definition of continuity, see page 1116 and 1099.) Equivalently, the reals are the minimal extension of the rationals forming a complete, totally strictly-Archimedean ordered field. This is the historical construction – or definition – of the integer, rational and real numbers from the natural numbers. However, it is not the only one construction possible. The most beautiful definition of all these types of numbers is the one discovered in 1969 by John Conway, and popularized by him, Donald Knuth and Martin Kruskal. Ref. 605

- A number is a sequence of bits. The two bits are usually called 'up' and 'down'. Examples of numbers and the way to write them are given in Figure 271.
- The empty sequence is zero.
- A finite sequence of n ups is the integer number n, and a finite sequence of n downs is the integer $-n$. Finite sequences of mixed ups and downs give the *dyadic rational numbers*. Examples are 1, 2, 3, −7, 19/4, 37/256, etc. They all have denominators with a power of 2. The other *rational numbers* are those that end in an infinitely repeating string of ups and downs, such as the *reals*, the *infinitesimals* and simple infinite numbers. Longer countably

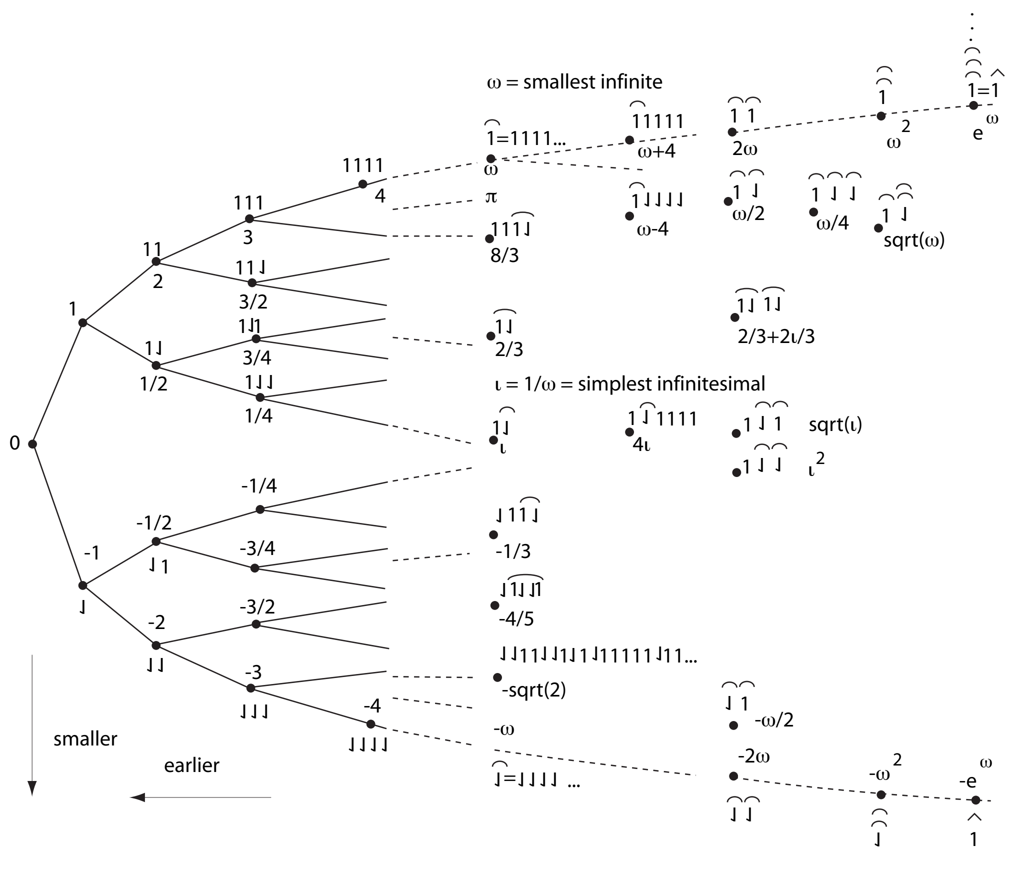


Figure 271 The surreal numbers in conventional and in bit notation

infinite series give even more crazy numbers. The complete class is called the class of *surreal* numbers.*

There is a second way to write surreal numbers. The first is the just mentioned sequence of bits. But in order to define addition and multiplication, another notation is usually used, deduced from Figure 271. A surreal α is defined as the earliest number of all those between two series of earlier surreals, the left and the right series:

$$\alpha = \{a, b, c, \ldots | A, B, C, \ldots\} \quad \text{with} \quad a, b, c, < \alpha < A, B, C \; . \tag{467}$$

For example, we have

$$\begin{gathered}\{0|\} = 1 \,,\ \{0,1|\} = 2 \,,\ \{|0\} = -1 \,,\ \{|-1,0\} = -2 \,,\ \{0\,|1\} = 1/2 \,, \\ \{0\,|1/2, 1/4\} = 1 \,,\ \{0,1,3/2,25/16\,|\,41/16,13/8,7/4,2\} = 1 + 37/64 \,,\end{gathered} \tag{468}$$

* The surreal numbers do *not* form a set since they contain all *ordinal numbers*, which themselves do not form a set, even though they of course *contain* sets. In short, ordinals and surreals are classes which are *larger* than sets.

showing that the finite surreals are the *dyadic numbers* $m/2^n$ (n and m being integers). Given two surreals $\alpha = \{..., a, ...|..., A, ...\}$ with $a < \alpha < A$ and $\beta = \{..., b, ...|..., B, ...\}$ with $b < \beta < B$, addition is defined recursively, using earlier, already defined numbers, as

$$\alpha + \beta = \{..., a + \beta, ..., \alpha + b, ...|..., A + \beta, ..., \alpha + B, ...\} . \tag{469}$$

This definition is used simply because it gives the same results as usual addition for integers and reals. Can you confirm this? By the way, addition is not always commutative. Are you able to find the exceptions, and to find the definition for subtraction? Multiplication is also defined recursively, namely by the expression

Challenge 1088 n

$$\begin{aligned}\alpha\beta =\{&..., a\beta + \alpha b - ab, ..., A\beta + \alpha B - AB, ...|\\ &..., a\beta + \alpha B - aB, ..., A\beta + \alpha b - Ab, ...\} .\end{aligned} \tag{470}$$

These definitions allow one to write $\iota = 1/\omega$, and to talk about numbers such as $\sqrt{\omega}$, the square root of infinity, about $\omega + 4$, $\omega - 1$, 2ω, e^{ω} and about other strange numbers shown in Figure 271. However, the surreal numbers are not commonly used. More common is one of their subsets.

Ref. 605

The *real numbers* are those surreals whose length is not larger than infinity and that do not have periodic endings with a period of length 1. In other words, the surreals distinguish the number $0.999999\overline{9}$ from the number 1, whereas the reals do not. In fact, between the two, there are infinitely many surreal numbers. Can you name a few?

Challenge 1089 n

Reals are more useful for describing nature than surreals, first because they form a set – which the surreals do not – and secondly because they allow the definition of integration. Other numbers defined with the help of reals, e.g. the complex numbers C and the quaternions H, are presented in Appendix D. A few more elaborate number systems are also presented there.

Appendix D

To conclude, in physics it is usual to call *numbers* the elements of any set that is a semiring (e.g. **N**), a ring (e.g. **Z**) or a field (**Q**, **R**, **C** or **H**). All these concepts are defined in Appendix D. Since numbers allow one to compare magnitudes and thus to measure, they play a central role in the description of observations.

Ref. 606

A series of equal balls is packed in such a way that the area of needed wrapping paper is minimal. For small numbers of balls the linear package, with all balls in one row, is the most efficient. For which number of balls is the linear package no longer a minimum?

Challenge 1090 n

Why use mathematics?

Die Forderung der Möglichkeit der einfachen Zeichen ist die Forderung der Bestimmtheit des Sinnes.*

Ludwig Wittgenstein, *Tractatus*, 3.23

Several well-known physicists have repeatedly asked why mathematics is so important.

* The requirement that simple signs be possible is the requirement that sense be determinate.

Ref. 607 For example, Niels Bohr is quoted as having said: 'We do not know why the language of mathematics has been so effective in formulating those laws in their most succinct form.' Eugene Wigner wrote an often cited paper entitled *The unreasonable effectiveness of mathematics.* Ref. 608 At the start of science, many centuries earlier, Pythagoras and his contemporaries were so overwhelmed by the usefulness of numbers in describing nature, that Pythagoras was able to organize a sect based on this connection. The members of the inner circle of this sect were called 'learned people,' in Greek '*mathematicians*', from the Greek μάθημα 'teaching'. This sect title then became the name of the modern profession.

These men forgot that numbers, as well as a large part of mathematics, are concepts developed precisely with the aim of describing nature. Numbers and mathematical concepts were developed right from the start to provide as succinct a description as possible. That is one consequence of mathematics being the science of symbolic necessities.

Perhaps we are being too dismissive. Perhaps these thinkers mainly wanted to express their feeling of wonder when experiencing that language works, that thinking and our brain works, and that life and nature are so beautiful. This would put the title question nearer to the well-known statement by Albert Einstein: 'The most incomprehensible fact about the universe is that it is comprehensible.' Comprehension is another word for description, i.e. for classification. Obviously, any separable system is comprehensible, and there is nothing strange about it. But is the universe separable? As long as is it described as being made of particles and vacuum, this is the case.

We will find in the third part of this adventure that the basic assumption made at our start is built on sand. The assumption that observations in nature can be *counted*, and thus that nature is separable, is an approximation. The quoted 'incomprehensibility' becomes amazement at the precision of this approximation. Nevertheless, Pythagoras' sect, which was based on the thought that 'everything in nature is numbers', was wrong. Like so many beliefs, observation will show that it was wrong.

> Die Physik ist für Physiker viel zu schwer.*
> David Hilbert (1862–1943), mathematician.

Is mathematics a language?

> Die Sätze der Mathematik sind Gleichungen, also Scheinsätze. Der Satz der Mathematik drückt keinen Gedanken aus.**
> Ludwig Wittgenstein, *Tractatus*, 6.2, 6.21

Surely, mathematics is a *vocabulary* that helps us to talk with precision. Mathematics can be seen as the exploration of *all* possible concepts that can be constructed from the two fundamental bricks 'set' and 'relation' (or some alternative, but equivalent pair). *Mathematics* is the science of symbolic necessities. Rephrased again, mathematics is the exploration of all possible types of classifications. This explains its usefulness in all situations

* Physics is much too difficult for physicists.

** The propositions of mathematics are equations, and therefore pseudo-propositions. A proposition of mathematics does not express a thought.

where complex, yet precise classifications of observations are necessary, such as in physics.

However, mathematics cannot express everything that humans want to communicate, such as wishes, ideas or feelings. Just try to express the fun of swimming using mathematics. Indeed, *mathematics is the science of symbolic necessities*; thus mathematics is not a language, nor does it contain one. Mathematical concepts, being based on *abstract* sets and relations, do not pertain to nature. Despite its beauty, mathematics does not allow us to talk about nature or the observation of motion. Mathematics does not tell *what* to say about nature; it does tell us *how* to say it.

In his famous 1900 lecture in Paris, the German mathematician David Hilbert* gave a list of 23 great challenges facing mathematics. The sixth of Hilbert's problems was to find a mathematical treatment of the axioms of physics. Our adventure so far has shown that physics started with *circular* definitions that has not yet been eliminated after 2500 years of investigations: space-time is defined with the help of objects and objects are defined Page 144 with the help of space and time. Being based on a circular definition, physics is thus *not* modelled after mathematics, even if many physicists and mathematicians, including Hilbert, would like it to be so. Physicists have to live with logical problems and have to walk on unsure ground in order to achieve progress. In fact, they have done so for 2500 years. If physics were an axiomatic system, it would not contain contradictions; on the other hand, it would cease to be a language and would cease to describe nature. We will Page 968 return to this issue later.

Curiosities and fun challenges

Challenge 1091 n ▪ What is the largest number that can be written with four digits of 2 and no other sign? And with four 4s?

▪ Pythagorean triplets are integers that obey $a^2 + b^2 = c^2$. Give at least ten examples.
Challenge 1092 e Then show the following three properties: at least one number in a triplet is a multiple of 3; at least one number in a triplet is a multiple of 4; at least one number in a triplet is a multiple of 5.

▪ The number $1/n$, when written in decimal notation, has a periodic sequence of digits. The period is at most $n-1$ digits long, as for $1/7 = 0.142857\,142857\,1428\ldots$. Which other
Challenge 1093 d numbers $1/n$ have periods of length $n-1$?

▪ Felix Klein was a famous professor of mathematics at Göttingen University. There were two types of mathematicians in his department: those who did research on whatever they wanted and those for which Klein provided the topic of research. To which type did
Challenge 1094 n Klein belong?

* David Hilbert (1862 Königsberg–1943 Göttingen), professor of mathematics in Göttingen, greatest mathematician of his time. He was a central figure to many parts of mathematics, and also played an important role both in the birth of general relativity and of quantum theory. His textbooks are still in print. His famous personal credo was: 'Wir müssen wissen, wir werden wissen.' (We must know, we will know.) His famous Paris lecture is published e.g. in *Die Hilbertschen Probleme*, Akademische Verlagsgesellschaft Geest & Portig, 1983. The lecture galvanized all of mathematics. (Despite efforts and promises of similar fame, *nobody* in the world had a similar overview of mathematics that allowed him or her to repeat the feat in the year 2000.) In his last decade he suffered the persecution of the Nazi regime; the persecution eliminated Göttingen from the list of important science universities, without recovering its place up to this day.

Obviously, this is a variation of another famous puzzle. A barber shaves all those people who do not shave themselves. Does the barber shave himself? Challenge 1095 n

▪ Everybody knows what a *magic square* is: a square array of numbers, in the simplest case from 1 to 9, that are distributed in such a way that the sum of all rows, columns (and possibly all diagonals) give the same sum. Can you write down the simplest $3 \times 3 \times 3$ *magic cube*? Challenge 1096 n

▪ The digits 0 to 9 are found on keyboards in two different ways. Calculators and keyboards have the 7 at the top left, whereas telephones and automatic teller machines have the digit 1 at the top left. The two standards, respectively by the International Standards Organization (ISO) and by the International Telecommunication Union (ITU, formerly CCITT), evolved separately and have never managed to merge. Ref. 609

▪ Leonhard Euler in his notebooks sometimes wrote down equations like

$$1 + 2^2 + 2^4 + 2^6 + 2^8 + \dots = -\frac{1}{3} \, . \tag{471}$$

Challenge 1097 d Can this make sense?

Physical concepts, lies and patterns of nature

> Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.*
>
> Ludwig Wittgenstein, *Tractatus*, 5.6

> Der Satz ist ein Bild der Wirklichkeit. Der Satz ist ein Modell der Wirklichkeit, so wie wir sie uns denken.**
>
> Ludwig Wittgenstein, *Tractatus*, 4.01

In contrast to mathematics, physics does aim at being a language. Through the description of motion it aims to express *everything* observed and, in particular, all examples and possibilities of change.*** Like any language, physics consists of concepts and sentences. In order to be able to express everything, it must aim to use few words for a lot of facts.****

* The limits of my language are the limits of my world.

** A proposition is a picture of reality. A proposition is a model of reality as we imagine it.

*** All observations are about change or variation. The various types of change are studied by the various sciences; they are usually grouped in the three categories of *human sciences*, *formal sciences* and *natural sciences*. Among the latter, the oldest are astronomy and metallurgy. Then, with the increase of curiosity in early antiquity, came the natural science concerned with the topic of motion: *physics*. In the course of our walk it will become clear that this seemingly restrictive definition indeed covers the whole set of topics studied in physics. In particular it includes the more common definition of physics as the study of matter, its properties, its components and their interactions.

**** A particular, specific observation, i.e. a specific example of input shared by others, is called a *fact*, or in other contexts, an *event*. A striking and regularly observed fact is called a *phenomenon*, and a general observation made in many different situations is called a *(physical) principle*. (Often, when a concept is introduced that is used with other meaning in other fields, in this walk it is preceded by the qualifier 'physical' or 'mathematical' in parentheses.) Actions performed towards the aim of collecting observations are called *experiments*. The concept of experiment became established in the sixteenth century; in the evolution of a child, it can best be compared to that activity that has the same aim of collecting experiences: *play*.

Physicists are essentially *lazy* people: they try to minimize the effort in everything they do. The concepts in use today have been optimised by the combined effort of many people to be as practical, i.e. as powerful as possible. A concept is called *powerful* when it allows one to express in a compact way a large amount of information, meaning that it can rapidly convey a large number of details about observations.

General statements about many examples of motion are called *rules* or *patterns*. In the past, it was often said that 'laws govern nature', using an old and inappropriate ideology. A physical 'law' is only a way of saying as much as possible with as few words as possible. When saying 'laws govern nature' we actually mean to say 'being lazy, we describe observations with patterns'. Laws are the epitome of laziness. Formulating laws is pure sloth. In fact, the correct expression is *patterns describe nature*.

Physicists have defined the laziness necessary for their field in much detail. In order to become a master of laziness, we need to distinguish lazy patterns from those which are not, such as lies, beliefs, statements that are not about observations, and statements that are not about motion. We do this below.

Page 615

The principle of extreme laziness is the origin, among others, of the use of numbers in physics. Observables are often best described with the help of numbers, because numbers allow easy and precise communication and classification. Length, velocity, angles, temperature, voltage or field strength are of this type. The notion of 'number', used in every measurement, is constructed, often unconsciously, from the notions of 'set' and 'relation', as shown above. Apart from the notion of number, other concepts are regularly defined to allow fast and compact communication of the 'laws' of nature; all are 'abbreviation tools.' In this sense, the statement 'the level of the Kac–Moody algebra of the Lagrangian of the heterotic superstring model is equal to one' contains precise information, explainable to everybody; however, it would take dozens of pages to express it using only the terms 'set' and 'relation.' In short, the *precision* common in physics results from its *quest for laziness*.

Are physical concepts discovered or created?

Das logische Bild der Tatsachen ist der Gedanke.*
Ludwig Wittgenstein, *Tractatus*, 3

The title question is often rephrased as: are physical concepts free of beliefs, taste or personal choices? The question has been discussed so much that it even appears in Hollywood movies. We give a short summary that can help you to distinguish honest from dishonest teachers.

Creation of concepts, in contrast to their discovery, would imply free choice between many alternative possibilities. The chosen alternative would then be due to the beliefs or tastes used. In physics (in obvious contrast to other, more ideological fields of enquiry), we know that different physical descriptions of observations are either equivalent or, in the opposite case, imprecise or even wrong. A description of observations is thus essentially unique: any choices of concepts are only apparent. There is no real freedom in the definition of physical concepts. In this property, physics is in strong contrast to artistic activity.

* A logical picture of facts is a thought.

If two different concepts can be used to describe the same aspect of observations, they must be equivalent, even if the relation that leads to the equivalence is not immediately clear. In fact, the requirement that people with different standpoints and observing the same event deduce equivalent descriptions lies at the very basis of physics. It expresses the requirement that observations are observer independent. In short, the strong requirement of viewpoint independence makes the free choice of concepts a logical impossibility.

The conclusion that concepts describing observations are discovered rather than created is also reached independently in the field of linguistics by the above-mentioned research on semantic primitives,* in the field of psychology by the observations on the formation of the concepts in the development of young children, and in the field of ethology by the observations of animal development, especially in the case of mammals. In all three fields detailed observations have been made of how the interactions between an individual and its environment lead to concepts, of which the most basic ones, such as space, time, object or interaction, are common across the sexes, cultures, races and across many animal species populating the world. Curiosity and the way that nature works leads to the same concepts for all people and even the animals; the world offers only one possibility, without room for imagination. Imagining that physical concepts can be created at your leisure is a belief – or a useful exercise, but never successful.

Physical concepts are classifications of observations. The activity of classification itself follows the patterns of nature; it is a mechanical process that machines can also perform. This means that any distinction, i.e. any statement that A is different from B, is a theory-free statement. No belief system is necessary to distinguish different entities in nature. Cats and pigs can also do so. Physicists *can* be replaced by animals, even by machines. Our mountain ascent will repeatedly confirm this point.

As already mentioned, the most popular physical concepts allow to describe observations as succinctly and as accurately as possible. They are formed with the aim of having the largest possible amount of understanding with the smallest possible amount of effort. Page 1030 Both Occam's razor – the requirement not to introduce unnecessary concepts – and the drive for unification automatically reduce the number and the type of concepts used in physics. In other words, the progress of physical science was and is based on a programme that reduces the possible choice of concepts as drastically as possible.

In summary, we found that physical concepts are the same for everybody and are free of beliefs and personal choices: they are first of all *boring*. Moreover, as they could stem from machines instead of people, they are *born of laziness*. Despite these human analogies – not meant to be taken too seriously – physical concepts are *not* created; they are discovered. If a teacher tells you the opposite, he is lying.

Having handled the case of physical concepts, let us now turn to physical statements. The situation is somewhat similar: physical statements must be lazy, arrogant and boring. Let us see why.

Wo der Glaube anfängt, hört die Wissenschaft auf.**
Ernst Haeckel, *Natürliche Schöpfungsgeschichte*, 1879.

* Anna Wierzbicka concludes that her research clearly indicates that semantic primitives are *discovered*, in Ref. 595 particular that they are deduced from the fundamentals of human experience, and not invented.
** Where belief starts, science ends.

Table 53 The 'scientific method'

Normal description **Curiosity**	Lobbyist description **Scientific method**
1. look around a lot	1. interact with the world
2. don't believe anything told	2. forget authority
3. choose something interesting and explore it yourself	3. observe
4. make up your own mind and describe precisely what you saw	4. use reason, build hypothesis
5. check if you can also describe similar situations in the same way	5. analyse hypothesis
6. increase the precision of observation until the checks either fail or are complete	6. perform experiments until hypothesis is proved false or established
7. depending on the case, continue with step 4 or 1	7. ask for more money

How do we find physical patterns and rules?

Grau, treuer Freund, ist alle Theorie,
Und grün des Lebens goldner Baum.*
J.W. v. Goethe, *Faust*.

Physics is usually presented as an objective science, but I notice that physics changes and the world stays the same, so there must be something subjective about physics.
Richard Bandler

Progressing through the study of motion reflects a young child's attitude towards life. The progress follows the simple programme on the left of Table 53.

Adult scientists do not have much more to add, except the more fashionable terms on the right, plus several specialized professions to make money from them. The experts of step 7 are variously called lobbyists or fund raisers; instead of calling this program 'curiosity', they call it the 'scientific method.' They mostly talk. Physics being the talk about motion,** and motion being a vast topic, many people specialize in this step.

The experts of step 6 are called *experimental physicists* or simply *experimentalists*, a term derived from the Latin 'experiri', meaning 'to try out'. Most of them are part of the

* 'Grey, dear friend, is all theory, and green the golden tree of life.' Johann Wolfgang von Goethe (1749–1832), the most influential German poet.

** Several sciences have the term 'talk' as part of their name, namely all those whose name finishes in '-logy', such as e.g. biology. The ending stems from ancient Greek and is deduced from λήγηιν meaning 'to say, to talk'. Physics as the science of motion could thus be called 'kinesiology' from κίνησις, meaning 'motion'; but for historical reasons this term has a different meaning, namely the study of human muscular activity. The term 'physics' is either derived from the Greek φύσικη (τέχνη is understood) meaning '(the art of) nature', or from the title of Aristotle' works τά φυσικά meaning 'natural things'. Both expressions are derived from φύσις, meaning 'nature'.

category 'graduate students'. The experts of steps 5 and 4 are called *theoretical physicists* or simply *theoreticians*.* This is a rather modern term; for example, the first professors of theoretical physics were appointed around the start of the twentieth century. The term is derived from the Greek θεωρία meaning 'observation, contemplation'. Finally, there are the people who focus on steps 1 to 3, and who induce others to work on steps 4 to 6; they are called *geniuses*.

Obviously an important point is hidden in step 6: how do all these people know whether their checks fail? How do they recognize truth?

> All professions are conspiracies against laymen.
>
> George Bernard Shaw

What is a lie?

> Get your facts straight, and then you can distort them at your leisure.
>
> Mark Twain (1835–1910)

> The pure truth is always a lie.
>
> Bert Hellinger

Lies are useful statements, as everybody learns during their youth. One reason that they are useful is because we can draw any imaginable conclusion from them. A well-known discussion between two Cambridge professors early in the twentieth century makes the point. McTaggart asked: 'If $2 + 2 = 5$, how can you prove that I am the pope?' Godfrey Hardy: 'If $2+2 = 5$, then $4 = 5$; subtract 3; then $1 = 2$; but McTaggart and the pope are two; therefore McTaggart and the pope are one.' As noted long ago, ex falso quodlibet. From what is wrong, anything imaginable can be deduced. It is true that in our mountain ascent we need to build on previously deduced results and that our trip could not be completed if we had a false statement somewhere in our chain of arguments. But lying is such an important activity that one should learn to perform it well.

There are various stages in the art of lying. Many animals have been shown to deceive their kin. Children start lying just before their third birthday, by hiding experiences. Ref. 580 Adults cheat on taxes. And many intellectuals or politicians even claim that truth does not exist.

However, in most countries, everybody must know what 'truth' is, since in a law court for example, telling an untruth can lead to a prison sentence. The courts are full of experts in lie detection. If you lie in court, you better do it well; experience shows that you might get away with many criminal activities. In court, a *lie* is a statement that contrasts with observations.* The truth of a statement is thus checked by observation. The check itself is

* If you like theoretical physics, have a look at the refreshingly candid web page by Nobel prize winner Gerard 't Hooft with the title *How to become a good theoretical physicist*. It can be found at http://www.phys.uu.nl/~thooft/theorist.html.

* Statements not yet checked are variously called *speculations*, *conjectures*, *hypotheses*, or – wrongly – simply *theses*. Statements that are in correspondence with observations are called *correct* or *true*; statements that contrast with observations are called *wrong* or *false*.

sometimes called the *proof* of the statement. For law courts, as for physics, *truth* is thus the correspondence with facts, and *facts* are shared observations. A 'good' lie is thus a lie whose contrast with shared observations is hard to discover.

The first way of lying is to put an emphasis on the sharedness only. Populists and polemics do this regularly. ('Every foreigner is a danger for the values of our country.') Since almost any imaginable opinion, however weird, is held by some group – and thus shared – one can always claim it as true. Unfortunately, it is no secret that ideas also get shared because they are fashionable, imposed or opposed to somebody who is generally disliked. Often a sibling in a family has this role – remember Cassandra.* For a good lie we thus need more than sharedness, more than *intersubjectivity* alone.

A good lie should be, like a true statement, really independent of the listener and the observer and, in particular, independent of their age, their sex, their education, their civilization or the group to which they belong. For example, it is especially hard – but not impossible – to lie with mathematics. The reason is that the basic concepts of mathematics, be they 'set', 'relation' or 'number', are taken from observation and are intersubjective, so that statements about them are easily checked. Usually, lies thus avoid mathematics.

Secondly, a 'good' lie should avoid statements about observations and use *interpretations* instead. For example, some people like to talk about other universes, which implies talking about fantasies, not about observations. A good lie has to avoid, however, to fall in the opposite extreme, namely to make statements which are meaningless; the most destructive comment that can be made about a statement is the one used by the great
Page 699 Austrian physicist Wolfgang Pauli: 'That is not even wrong.'

Thirdly, a good lie doesn't care about observations, only about imagination. Only truth needs to be *empirical*, to distinguish it from *speculative* statements. If you want to lie 'well' even with empirical statements, you need to pay attention. There are two types of empirical statements: *specific* statements and *universal* statements. For example, 'On the 31st of August 1960 I saw a green swan swimming on the northern shore of the lake of Varese' is specific, whereas 'All ravens are black' is universal, since it contains the term 'all'. There is a well-known difference between the two, which is important for lying well: specific statements cannot be falsified, they are only verifiable, and universal statements
Ref. 610 cannot be verified, they are only falsifiable. Why is this so?

Universal statements such as 'the speed of light is constant' cannot be tested for *all* possible cases. (Note that if they could, they would not be universal statements, but just a list of specific ones.) However, they can be reversed by a counterexample. Another example of the universal type is: 'Apples fall upwards.' Since it is falsified by an observation conducted by Newton several centuries ago, or by everyday experience, it qualifies as an (easily detectable) lie. In general therefore, lying by stating the opposite of a theory is usually unsuccessful. If somebody insists on doing so, the lie becomes a *superstition*, a *belief*, a *prejudice* or a *doctrine*. These are the low points in the art of lying. A famous case of insistence on a lie is that of the colleagues of Galileo, who are said to have refused to

* The implications of birth order on creativity in science and on acceptance of new ideas has been studied in the fascinating book by Frank J. Sulloway, *Born to Rebel – Birth Order, Family Dynamics and Creative Lives*, Panthon Books, 1996. This exceptional book tells the result of a life-long study correlating the personal situations in the families of thousands of people and their receptivity to about twenty revolutions in the recent history. The book also includes a test in which the reader can deduce their own propensity to rebel, on a scale from 0 to 100 %. Darwin scores 96 % on this scale.

look through his telescope to be convinced that Jupiter has moons, an observation that would have shaken their belief that everything turns around the Earth. Obviously these astronomers were amateurs in the art of lying. A good universal lie is one whose counter-example is not so easily spotted.

There should be no insistence on lies in physics. Unfortunately, classical physics is full of lies. We will dispel them during the rest of our walk.

Lying by giving specific instead of universal statements is much easier. ('I can't remember.') Even a specific statement such as 'yesterday the Moon was green, cubic and smelled of cheese' can never be completely falsified: there is no way to show with absolute certainty that this is wrong. The only thing that we can do is to check whether the statement is compatible with other observations, such as whether the different shape affected the tides as expected, whether the smell can be found in air collected that day, etc. A good specific lie is thus not in contrast with other observations.*

Incidentally, universal and specific statements are connected: the *opposite* of a universal statement is always a specific statement, and vice versa. For example, the opposite of the general statement 'apples fall upwards', namely 'some apples fall downwards', is specific. Similarly, the the specific statement 'the Moon is made of green cheese' is in opposition to the universal statement 'the Moon is solid since millions of years and has almost no smell or atmosphere.'

In other words, law courts and philosophers disagree. Law courts have no problem with calling theories true, and specific statements lies. Many philosophers avoid this. For example, the statement 'ill-tempered gaseous vertebrates do not exist' is a statement of the universal type. If a universal statement is in agreement with observations, and if it is falsifiable, law courts call it *true*. The opposite, namely the statement: 'ill-tempered gaseous vertebrates do exist', is of the *specific* type, since it means 'Person X has observed an ill-tempered gaseous vertebrate in some place Y at some time Z'. To verify this, we need a record of the event. If such a record, for example a photographs or testimony does not exist, and if the statement *can* be falsified by other observations, law courts call the specific statement a *lie*. Even though these are the rules for everyday life and for the law, there is no agreement among philosophers and scientists that this is acceptable. Why? Intellectuals are a careful lot, because many of them have lost their lives as a result of exposing lies too openly.

In short, specific lies, like all specific statements, can never be falsified with certainty. This is what makes them so popular. Children learn specific lies first. ('I haven't eaten the jam.') General lies, like all general statements, can always be corroborated by examples.

* It is often difficult or tedious to verify statements concerning the past, and the difficulty increases with the distance in time. That is why people can insist on the occurrence of events which are supposed to be exceptions to the patterns of nature ('miracles'). Since the advent of rapid means of communication these checks are becoming increasingly easy, and no miracles are left over. This can be seen in Lourdes in France, where even though today the number of visitors is much higher than in the past, no miracles have been seen in decades.

In fact, all modern 'miracles' are kept alive only by consciously eschewing checks, such as the supposed yearly liquefaction of blood in Napoli, the milk supposedly drunk by statues, the supposed healers in television evangelism, etc. Most miracles only remain because many organizations make money out of the difficulty of falsifying specific statements. For example, when the British princess Diana died in a car crash in 1997, even though the events were investigated in extreme detail, the scandal press could go on almost without end about the 'mysteries' of the accident.

Ref. 611

This is the reason for the success of ideologies. But the criteria for recognizing lies, even general lies, have become so commonplace that beliefs and lies try to keep up with them. It became fashionable to use expressions such as 'scientific fact' – there are no non-scientific facts –, or 'scientifically proven' – observations cannot be proven otherwise – and similar empty phrases. These are not 'good' lies; whenever we encounter sentences beginning with 'science says ...' or 'science and religion do ...', replacing 'science' by 'knowledge' or 'experience' is an efficient way of checking whether such statements are to be taken seriously or not.*

Lies differ from true statements in their emotional aspect. Specific statements are usually boring and fragile, whereas specific lies are often sensational and violent. In contrast, general statements are often daring and fragile whereas general lies are usually boring and violent. The truth is fragile. True statements require the author to stick his neck out to criticism. Researchers know that if one doesn't stick the neck out, it can't be an observation or a theory. (A *theory* is another name for one or several connected, not yet falsified universal statements.)* Telling the truth does make vulnerable. For this reason, theories are often *daring*, *arrogant* or *provoking*; at the same time they have to be *fragile* and *vulnerable*. For men, theories thus resemble what they think about women. Darwin's *The origin of the species*, which developed daring theories, illustrates the stark contrast between the numerous boring and solid facts that Darwin collected and the daring theory that he deduced. Boredom of facts is a sign of truth.

In contrast, the witch-hunters propagating the so-called 'intelligent design' are examples of liars. The specific lies they propagate, such as 'the world was created in October 4004 bce', are sensational, whereas the general lies they propagate, such as 'there have not been big changes in the past', are boring. This is in full contrast with common sense. Moreover, lies, in contrast to true statements, make people violent. The worse the lie, the more violent the people. This connection can be observed regularly in the news. In other words, 'intelligent design' is not only a lie, it is a bad lie. A 'good' *general lie*, like a good physical theory, seems crazy and seems vulnerable, such as 'people have free will'. A 'good' specific lie is boring, such as 'this looks like bread, but for the next ten minutes it is not'. Good lies do not induce violence. Feelings can thus be a criterion to judge the quality of lies, if we pay careful attention to the type of statement. A number of common lies are discussed later in this intermezzo.

An important aspect of any 'good' lie is to make as few *public* statements as possible, so that critics can check as little as possible. (For anybody sending corrections of mistakes in this text, the author provides a small reward.) To detect lies, public scrutiny is important,

* To clarify the vocabulary usage of this text: *religion* is spirituality plus a varying degree of power abuse. The mixture depends on each person's history, background and environment. *Spirituality* is the open participation in the whole of nature. Most, maybe all, people with a passion for physics are spiritual. Most are not religious.

* In other words, a set of not yet falsified patterns of observations on the same topic is called a *(physical) theory*. The term 'theory' will always be used in this sense in this walk, i.e. with the meaning 'set of correct general statements'. This use results from its Greek origin: 'theoria' means 'observation'; its original meaning, 'passionate and emphatic contemplation', summarizes the whole of physics in a single word. ('Theory', like 'theatre', is formed from the root θέ, meaning 'the act of contemplating'.) Sometimes, however, the term 'theory' is used – being confused with 'hypothesis' – with the meaning of 'conjecture', as in 'your theory is wrong', sometimes with the meaning of 'model', as in 'Chern–Simons' theory and sometimes with the meaning of 'standard procedure', as in 'perturbation theory'. These incorrect uses are avoided here.

though not always reliable. Sometimes, even scientists make statements which are not based on observations. However, a 'good' lie is always well prepared and told on purpose; accidental lies are frowned upon by experts. Examples of good lies in science are 'aether', 'UFOs', 'creation science', or 'cold fusion'. Sometimes it took many decades to detect the lies in these domains.

To sum up, the central point of the art of lying without being caught is simple: do not divulge details. Be *vague*. All the methods used to verify a statement ask for details, for *precision*. For any statement, its degree of precision allows one to gauge the degree to which the author is sticking his neck out. The more precision that is demanded, the weaker a statement becomes, and the more likely a fault will be found, if there is one. This is the main reason that we chose an increase in precision as a guide for our mountain ascent. By the way, the same method is used in criminal trials. To discover the truth, investigators typically ask all the witnesses a large number of questions, allowing as many *details* as possible come to light. When sufficient details are collected, and the precision is high enough, the situation becomes clear. Telling 'good' lies is much more difficult than telling the truth; it requires an excellent imagination.

> Truth is an abyss.
>
> Democritus

> To teach superstitions as truth is a most terrible thing.
>
> Hypatia of Alexandria (*c.* 355–415)

> [Absolute truth:] It is what scientists say it is when they come to the end of their labors.
>
> Charles Peirce (1839–1914)

Ref. 613

Is this statement true?

> Truth is a rhetorical concept.
>
> Paul Feyerabend (b. 1924 Vienna, d. 1994 Zürich)

Not all statements can be categorized as true or false. Statements can simply make no sense. There are even such statements in mathematics, where they are called *undecidable*. An example is the continuum hypothesis. This hypothesis is undecidable because it makes a statement that depends on the precise meaning of the term 'set'; in standard mathematical usage the term is not defined sufficiently precisely so that a truth value can be assigned to the continuum hypothesis. In short, statements can be undecidable because the concepts contained in them are not sharply defined.

Statements can also be undecidable for other reasons. Phrases such as 'This statement is not true' illustrate the situation. Kurt Gödel* has even devised a general way of constructing such statements in the domain of logic and mathematics. The different variations of these *self-referential* statements, especially popular both in the field of logic and

computer science, have captured a large public.* Similarly undecidable statements can be
Ref. 614 constructed with terms such as 'calculable', 'provable' and 'deducible'.

In fact, self-referential statements are undecidable because they are meaningless. If the usual definition of 'true', namely corresponding to facts, is substituted into the sentence 'This statement is not true', we quickly see that it has no meaningful content. The most famous meaningless sentence of them all was constructed by the linguist Noam Chomsky:

Colorless green ideas sleep furiously.

Ref. 586 It is often used as an example for the language processing properties of the brain, but nobody sensible elevates it to the status of a paradox and writes philosophical discussions about it. To do that with the title of this section is a similar waste of energy.

The main reason for the popular success of self-reference is the difficulty in perceiving the lack of meaning.* A good example is the statement:

This statement is false or you are an angel.

Challenge 1099 n We can actually deduce from it that 'you are an angel.' Can you see how? If you want, you can change the second half and get even more interesting statements. Such examples show that statements referring to themselves have to be treated with great care when under investigation.

In physics, in the other natural sciences and in legal trials these problems do not emerge, since self-referential statements are not used. In fact, the work of logicians confirms, often rather spectacularly, that there is no way to extend the term 'truth' beyond the definition of 'correspondence with facts.'

> Ein Satz kann unmöglich von sich selbst aussagen, daß er wahr ist.**
>
> Ludwig Wittgenstein, *Tractatus*, 4.442

Challenges about lies

Some lies are entertaining, others are made with criminal intent; some are good, others are bad.

Challenge 1100 e ▪ 'Yesterday I drowned.' Is this a good or a bad lie?

* Kurt Gödel (1906–1978), famous Austrian logician.

* A general introduction is given in the beautiful books by Raymond Smullyan: *Satan, Cantor and Infinity and Other Mind-boggling Puzzles*, Knopf, 1992; *What is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles*, Touchstone, 1986, and *The Lady or the Tiger? And Other Puzzles*, Times Books, 1982. Also definitions can have no content, such as David Hilbert's 'smallest number that has not been mentioned this century' or 'the smallest sequence of numbers that is described by more signs than this sentence'.

* A well-known victim of this difficulty is Paulus of Tarsus. The paradox of the Cretan poet *Epimenedes* (6th century BCE) who said 'All Cretans lie' is too difficult for the notoriously humour-impaired Paulus, who in his letter to Titus (chapter 1, verses 12 and 13, in the christian bible) calls Epimenedes a 'prophet',
Ref. 615 adds some racist comments, and states that this 'testimony' is true. But wait; there is a final twist to this story. The statement 'All Cretans lie' is *not* a paradox at all; a truth value can actually be ascribed to it, because the
Challenge 1098 n statement is not really self-referential. Can you confirm this? The only *genuine* paradox is 'I am lying', to which it is indeed impossible to ascribe a truth value.

** It is quite impossible for a proposition to state that it itself is true.

▪ In the 1990s, so-called *crop circles* were formed by people walking with stilts, a piece of wood and some rope in fields of crops. Nevertheless, many said and others believed that these circles were made by extraterrestrial beings. Is this a good or a bad lie? Can you give a few reasons why this is impossible?

Challenge 1101 ny

▪ Sometimes it is heard that a person whose skin is completely covered with finest metal powder will die, due to the impossibility of the skin to breathe. Can you show that this is wrong?

Challenge 1102 ny

▪ A famous hoax premises that the Earth is only about six thousand years old. (Believers regularly use this lie as justification for violence against non-believers.) Can you explain why this is wrong?

Challenge 1103 ny

▪ A famous provocation: the world has been created last Saturday. Can you decide whether this is wrong?

Challenge 1104 ny

▪ Hundreds of hoaxes are found on the http://www.museumofhoaxes.com web site. It gives an excellent introduction into the art of lying; of course it exposes only those who have been caught. Enjoy the science stories, especially those about archaeology. (Several other sites with similar content can be found on the internet.)

▪ In the 1990s, many so-called 'healers' in the Philippines made millions by suggesting patients that they were able to extract objects from their bodies without operating. Why is this not possible? (For more information on health lies, see the http://www.quackwatch.com website.)

Challenge 1105 e

▪ Since the 1980s, people have claimed that it is possible to acquire knowledge simply from somebody 1000 km away, without any communication between the two people. However, the assumed 'morphogenetic fields' cannot exist. Why not?

Challenge 1106 e

▪ It is claimed that a Fire Brigade building in a city in the US hosts a light bulb that has been burning without interruption since 1901 (at least it was so in 2005). Can this be true? Hundreds of such stories, often called 'urban legends,' can be found on the http://www.snopes.com website. However, some of the stories are not urban legends, but true, as the site shows.

Challenge 1107 n

▪ 'This statement has been translated from French into English'. Is the statement true, false or neither?

Observations

> Knowledge is a sophisticated statement of ignorance.
>
> Attributed to Karl Popper

The collection of a large number of true statements about a type of observations, i.e. of a large number of facts, is called *knowledge*. Where the domain of observations is sufficiently extended, one speaks of a *science*. A *scientist* is thus somebody who collects knowledge.* We found above that an observation is classified input into the memory of several

Ref. 616

* The term 'scientist' is a misnomer peculiar to the English language. Properly speaking, a 'scientist' is a follower of *scientism*, an extremist philosophical school that tried to resolve all problems through science. For this reason, some religious sects have the term in their name. Since the English language did not have a shorter term to designate 'scientific persons', as they used to be called, the term 'scientist' started to appear in the United States, from the eighteenth century onwards. Nowadays the term is used in all English-speaking countries – but not outside them, fortunately.

people. Since there is motion all around, to describe all these observations is a mammoth task. As for every large task,to a large extent the use of appropriate tools determines the degree of success that can be achieved. These tools, in physics and in all other sciences, fall in three groups: tools for the collection of observations, tools to communicate observations and tools to communicate relations between observations. The latter group has been already discussed in the section on language and on mathematics. We just touch on the other two.

Have enough observations been recorded?

> Every generation is inclined to define 'the end of physics' as coincident with the end of their scientific contributions.
>
> Julian Schwinger*

Physics is an experimental science; it rests on the collection of observations. To realize this task effectively, all sorts of *instruments*, i.e. tools that facilitate observations, have been developed and built. Microscopes, telescopes, oscilloscopes, as well as thermometers, hygrometers, manometers, pyrometers, spectrometers amongst others are familiar examples. The precision of many of these tools is being continuously improved even today; their production is a sizeable part of modern industrial activity, examples being electrical measuring apparatus and diagnostic tools for medicine, chemistry and biology. Instruments can be as small as a tip of a few tungsten atoms to produce an electron beam of a few volts, and as large as 27 km in circumference, producing an electron beam with more than 100 GV effective accelerating voltage. Instruments have been built that contain and measure the coldest known matter in the universe. Other instruments can measure length variations of much less than a proton diameter over kilometre long distances. Instruments have been put deep inside the Earth, on the Moon, on several planets, and have been sent outside the Solar system.

In this walk, instruments are not described; many good textbooks on this topic are available. Ref. 618, Ref. 619 Most observations collected by instruments are not mentioned here. The most important results in physics are recorded in standard publications, such as the Landolt–Börnstein series and the physics journals Ref. 621 (Appendix E gives a general overview of information sources).

Will there be significant new future observations in the domain of the fundamentals of motion? At present, *in this specific domain*, even though the number of physicists and publications is at an all-time high, the number of new experimental discoveries has been steadily diminishing for many years and is now fairly small. The sophistication and investment necessary to obtain new results has become extremely high. In many cases, measuring instruments have reached the limits of technology, of budgets or even those of nature. The number of new experiments that produce results showing no deviation

* Julian Seymour Schwinger (1918–1994), US-American infant prodigy. He was famous for his clear thinking and his excellent lectures. He worked on waveguides and synchroton radiation, made contributions to nuclear physics and developed quantum electrodynamics. For the latter he received the 1965 Nobel prize in physics together with Tomonaga and Feynman. He was a thesis advisor to many famous physicists and Ref. 617 wrote several excellent and influential textbooks. At the end of his life, he became strangely interested in cold fusion.

from theoretical predictions is increasing steadily. The number of historical papers that try to enliven dull or stalled fields of enquiry are increasing. Claims of new effects which turn out to be false, due to measurement errors, self-deceit or even fraud have become so frequent that scepticism has become a common response. Although in many domains of science, including physics, discoveries are still expected, on the fundamentals of motion the arguments just presented seem to show that new observations are only a remote possibility. The task of collecting observations on the foundations of motion (though not on other topics of physics) seems to be *complete*. Indeed, most observations described here were obtained before the end of the twentieth century. We are not too early with our walk.

> Measure what is measurable; make measurable what is not.
>
> Wrongly attributed to Galileo.

Ref. 620

Are all physical observables known?

> Scientists have odious manners, except when you prop up their theory; then you can borrow money from them.
>
> Mark Twain (1835–1910)

The most practical way to communicate observations was developed a long time ago: by measurements. A measurement allows effective communication of an observation to other times and places. This is not always as trivial as it sounds; for example, in the middle ages people were unable to compare precisely the 'coldness' of the winters of two different years! The invention of the thermometer provided a reliable solution to this requirement. A *measurement* is thus the classification of an observation into a standard set of observations; to put it simply, a measurement is a *comparison with a standard*. This definition of a measurement is precise and practical, and has therefore been universally adopted. For example, when the length of a house is measured, this aspect of the house is classified into a certain set of standard lengths, namely the set of lengths defined by multiples of a unit. A *unit* is the abstract name of the standard for a certain observable. Numbers and units allow the most precise and most effective communication of measurement results.

For all measurable quantities, practical standard units and measurement methods have been defined; the main ones are listed and defined in Appendix B. All units are derived from a few fundamental ones; this is ultimately due to our limited number of senses: length, time and mass are related to sight, hearing and touch. Our limited number of senses is, in turn, due to the small number of observables of nature.

We call *observables* the different measurable aspects of a system. Most observables, such as size, speed, position, etc. can be described by numbers, and in this case they are *quantities*, i.e. multiples of some standard unit. Observables are usually abbreviated by *(mathematical) symbols*, usually letters from some alphabet. For example, the symbol c commonly specifies the velocity of light. For most observables, standard symbols have been defined by international bodies.* The symbols for the observables that describe the state of an object are also called *variables*. Variables on which other observables depend are often called *parameters*. (Remember: a parameter is a variable constant.) For example, the speed of light is a constant, the position a variable, the temperature is often a parameter, on which the length of an object, for example, can depend. Note that not all observables are quantities; in particular, parities are not multiples of any unit.

Today the task of defining tools for the communication of observations can be considered *complete*. (For quantities, this is surely correct; for parity-type observables there could be a few examples to be discovered.) This is a simple and strong statement. Even the BIPM, the Bureau International des Poids et Mesures, has stopped adding new units.*

As a note, the greatness of a physicist can be ranked by the number of observables he has introduced. Even a great scientist such as Einstein, who discovered many 'laws' of nature, only introduced one new observable, namely the metric tensor for the description of gravity. Following this criterion – as well as several others – Maxwell is the most important physicist, having introduced electric and magnetic fields, the vector potential, and several other material dependent observables. For Heisenberg, Dirac and Schrödinger, the wavefunction describing electron motion could be counted as half an observable (as it is a quantity necessary to calculate measurement results, but not itself an observable). Incidentally, even the introduction of *any* term that is taken up by others is a rare event; 'gas', 'entropy' and only a few others are such examples. It has always been much more difficult to discover an observable than to discover a 'law'; usually, observables are developed by many people cooperating together. Indeed, many 'laws' bear people's names, but almost no observables.

If the list of observables necessary to describe nature is complete, does this mean that all the patterns or rules of nature are known? No; in the history of physics, observables were usually defined and measured long *before* the precise rules connecting them were found. For example, all observables used in the description of motion itself, such as time, position and its derivatives, momentum, energy and all the thermodynamic quantities, were defined before or during the nineteenth century, whereas the most precise versions of the patterns or 'laws' of nature connecting them, special relativity and non-equilibrium thermodynamics, have been found only in the twentieth century. The same is true for all observables connected to electromagnetic interaction. The corresponding patterns of nature, quantum electrodynamics, was discovered long after the corresponding observables. The observables that were discovered last were the fields of the strong and the weak nuclear interactions. Also, in this case, the patterns of nature were formulated much later.**

* All mathematical symbols used in this walk, together with the alphabets from which they are taken, are listed in Appendix A on notation. They follow international standards whenever they are defined. The standard symbols of the physical quantities, as defined by the International Standards Organization (ISO), the International Union of Pure and Applied Physics (IUPAP) and the International Union of Pure and Applied Chemistry (IUPAC), can be found for example in the *bible*, i.e. the *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, 1992.

* The last, the katal or mol/s, was introduced in 1999. Physical units are presented in Appendix B.

** Is it possible to talk about observations at all? It is many a philosopher's hobby to discuss whether there actually is an example for an 'Elementarsatz' mentioned by Wittgenstein in his *Tractatus*. There seems to be at least one that fits: *Differences exist*. It is a simple sentence; in the third part of our walk, it will play a central role.

Do observations take time?

An observation is an interaction with some part of nature leading to the production of a record, such as a memory in the brain, data on a tape, ink on paper, or any other fixed process applied to a support. The necessary irreversible interaction process is often called *writing* the record. Obviously, writing takes a certain amount of time; zero interaction time would give no record at all. Therefore any recording device, including our brain, always records some *time average* of the observation, however short it may be.

What we call a fixed image, be it a mental image or a photograph, is always the time average of a moving situation. Without time averaging, we would have no fixed memories. On the other hand, any time averaging introduces a blur that hides certain details; and in our quest for precision, at a certain moment, these details are bound to become important. The discovery of these details will begin in the second part of the walk, the one centred on quantum theory. In the third part of our mountain ascent we will discover that there is a shortest possible averaging time. Observations of that short duration show so many details that even the distinction between particles and empty space is lost. In contrast, our concepts of everyday life appear only after relatively long time averages. The search for an average-free description of nature is one of the big challenges of our adventure.

Is induction a problem in physics?

> Nur *gesetzmäßige* Zusammenhänge sind *denkbar*.*
>
> Ludwig Wittgenstein, *Tractatus*, 6.361

> There is a tradition of opposition between adherents of induction and of deduction. In my view it would be just as sensible for the two ends of a worm to quarrel.
>
> Alfred North Whitehead (1861–1947)

Induction is the usual term used for the act of making, from a small and finite number of experiments, general conclusions about the outcome of *all* possible experiments performed in other places, or at other times. In a sense, it is the technical term for sticking out one's neck, which is necessary in every scientific statement. Induction has been a major topic of discussion for science commentators. Frequently one finds the remark that knowledge in general, and physics in particular, relies on induction for its statements. According to some, induction is a type of hidden belief that underlies all sciences but at the same time contrasts with them.

To avoid wasting energy, we make only a few remarks. The first can be deduced from a simple experiment. Try to convince a critic of induction to put their hand into a fire. Nobody who honestly calls induction a belief should conclude from a few unfortunate experiences in the past that such an act would also be dangerous in the future... In short, somehow induction works.

A second point is that physical universal statements are always openly stated; they are never hidden. The refusal to put one's hand into a fire is a consequence of the invariance

* Only connexions that are *subject to law* are *thinkable*.

of observations under time and space translations. Indeed, general statements of this type form the very basis of physics. However, no physical statement is a belief only because it is universal; it always remains open to experimental checks. Physical induction is not a hidden method of argumentation, it is an explicit part of experimental statements. In fact, the complete list of 'inductive' statements used in physics is given in the table on page 184. These statements are so important that they have been given a special name: they are called *symmetries*. The table lists all known symmetries of nature; in other words, it lists all inductive statements used in physics.

Perhaps the best argument for the use of induction is that there is no way to avoid it when one is thinking. There is no way to think, to talk or to remember without using concepts, i.e. without assuming that most objects or entities have the same properties over time. There is also no way to communicate with others without assuming that the observations made from the other's viewpoint are similar to one's own. There is no way to think without symmetry and induction. Indeed, the concepts related to symmetry and induction, such as space and time, belong to the fundamental concepts of language. The only sentences which do not use induction, the sentences of logic, do not have any content (*Tractatus*, 6.11). Indeed, without induction, we cannot classify observations at all! Evolution has given us memory and a brain because induction works. To criticize induction is not to criticize natural sciences, it is to criticize the use of thought in general. We should never take too seriously people who themselves do what they criticize in others; sporadically pointing out the ridicule of this endeavour is just the right amount of attention they deserve.

Page 596
Ref. 616
Challenge 1108 n

The topic could be concluded here, were it not for some interesting developments in modern physics that put two additional nails in the coffin of arguments against induction. First, in physics whenever we make statements about all experiments, all times or all velocities, such statements are actually about a *finite number* of cases. We know today that infinities, both in size and in number, do not occur in nature. The infinite number of cases appearing in statements in classical physics and in quantum mechanics are apparent, not real, and due to human simplifications and approximations. Statements that a certain experiment gives the same result 'everywhere' or that a given equation is correct for 'all times', always encompass only a *finite* number of examples. A great deal of otherwise often instinctive repulsion to such statements is avoided in this way. In the sciences, as well as in this book, 'all' *never* means an infinite number of cases.

Finally, it is well known that extrapolating from a few cases to many is false when the few cases are independent of each other. However, this conclusion is correct if the cases are interdependent. From the fact that somebody found a penny on the street on two subsequent months, cannot follow that he will find one the coming month. Induction is only correct if we know that all cases have similar behaviour, e.g. because they follow from the same origin. For example, if a neighbour with a hole in his pocket carries his salary across that street once a month, and the hole always opens at that point because of the beginning of stairs, then the conclusion would be correct. It turns out that the results of modern physics encountered in the third part of our walk show that all situations in nature are indeed interdependent, and thus we prove in detail that what is called 'induction' is in fact a logically correct conclusion.

> In the progress of physics, the exception usually turned out to be the general case.

The quest for precision and its implications

> Der Zweck der Philosophie ist die logische Klärung der Gedanken.*
> Ludwig Wittgenstein, *Tractatus*, 4.112

To talk well about motion means to talk precisely. Precision requires avoiding three common mistakes in the description of nature.

First, concepts should never have a contradiction built into their definition. For example, any phenomenon occurring in nature evidently is a 'natural' phenomenon; therefore, to talk about either 'supernatural' phenomena or 'unnatural' phenomena is a mistake that nobody interested in motion should let go unchallenged; such terms contain a logical contradiction. Naturally, *all* observations are natural. Incidentally, there is a reward of more than a million dollars for anybody proving the opposite. In over twenty years, nobody has yet been able to collect it. Ref. 622

Second, concepts should not have unclear or constantly changing definitions. Their content and their limits must be kept constant and explicit. The opposite of this is often encountered in crackpots or populist politicians; it distinguishes them from more reliable thinkers. Physicists can also fall into the trap; for example, there is, of course, only one *single* (physical) universe, as even the name says. To talk about more than one universe is an increasingly frequent error. Ref. 623

Third, concepts should not be used outside their domain of application. It is easy to succumb to the temptation to transfer results from physics to philosophy without checking the content. An example is the question: 'Why do particles follow the laws of nature?' The flaw in the question is due to a misunderstanding of the term 'laws of nature' and to a confusion with the laws of the state. If nature were governed by 'laws', they could be changed by parliament. Remembering that 'laws of nature' simply means 'pattern', 'property' or 'description of behaviour', and rephrasing the question correctly as 'Why do particles behave in the way we describe their behaviour?' one can recognize its senselessness.

In the course of our walk, we will often be tempted by these three mistakes. A few such situations follow, with the ways of avoiding them.

> Consistency is the last refuge of the unimaginative.
> Oscar Wilde (b. 1854 Dublin, d. 1900 Paris)

* Philosophy aims at the logical clarification of thoughts.

What are interactions? – No emergence

The whole is always more than the sum of its parts.
Aristotle, *Metaphysica*, 10f–1045a.

In the physical description of nature, the whole is always *more* than the sum of its parts. Actually, the difference between the whole and the sum of its parts is so important that it has a special name: the *interaction* between the parts. For example, the energy of the whole minus the sum of the energies of its parts is called the energy of interaction. In fact, the study of interactions is the main topic of physics. In other words, physics is concerned *primarily* with the difference between the parts and the whole, contrary to what is often suggested by bad journalists or other sloppy thinkers.

Note that the term 'inter-action' is based on the general observation that anything that affects anything else is, in turn, affected by it; interactions are *reciprocal*. For example, if one body changes the momentum of another, then the second changes the momentum of the first by the same (negative) amount. The reciprocity of interactions is a result of conservation 'laws'. The reciprocity is also the reason that somebody who uses the term 'interaction' is considered a heretic by monotheistic religions, as theologians regularly point out. They repeatedly stress that such a reciprocity implicitly denies the immutability of the deity. (Are they correct?) Challenge 1109 ny

The application of the definition of interaction also settles the frequently heard question of whether in nature there are 'emergent' properties, i.e. properties of systems that cannot be deduced from the properties of their parts and interactions. By definition, there are no emergent properties. 'Emergent' properties can only appear if interactions are approximated or neglected. Page 237 The idea of 'emergent' properties is a product of minds with restricted horizons, unable to see or admit the richness of consequences that general principles can produce. Ref. 624 In defending the idea of emergence, one belittles the importance of interactions, working, in a seemingly innocuous, maybe unconscious, but in fact sneaky way, against the use of reason in the study of nature. 'Emergence' is a belief.

The simple definition of interaction given above sounds elementary, but it leads to surprising conclusions. Take the atomic idea of Democritus in its modern form: nature is made of vacuum and of particles. The first consequence is the *paradox of incomplete description*: experiments show that there are interactions between vacuum and particles. However, interactions are differences between parts and the whole, in this case between vacuum and particles on the one hand, and the whole on the other. We thus have deduced that nature is not made of vacuum and particles alone.

The second consequence is the *paradox of overcomplete description*: experiments also show that interactions happen through exchange of particles. However, we have counted particles already as basic building blocks. Does this mean that the description of nature by vacuum and particles is an overdescription, counting things twice? Challenge 1110 n

We will resolve both paradoxes in the third part of the mountain ascent. Page 939

What is existence?

You know what I like most?
Rhetorical questions.

Ref. 625 Assume a friend tells you 'I have seen a *grampus* today!' You would naturally ask what it looks like. What answer do we expect? We expect something like 'It's an animal with a certain number of heads similar to a X, attached to a body like a Y, with wings like a Z, it make noises like a U and it felt like a V' – the letters denoting some other animal or object. Generally speaking, in the case of an object, this scene from Darwin's voyage to South America shows that in order to talk to each other, we first need certain basic, common concepts ('animal', 'head', 'wing', etc.). In addition, for the definition of a new entity we need a characterization of its parts ('size', 'colour'), of the way these parts relate to each other, and of the way that the whole interacts with the outside world ('feel', 'sound'). In other words, for an object to exist, we must be able to give a list of relations with the outside world. An object exists if we can interact with it. (Is observation sufficient to determine existence?) Challenge 1111 n

For an abstract concept, such as 'time' or 'superstring', the definition of existence has to be refined only marginally: *(physical) existence is the effectiveness to describe interactions accurately*. This definition applies to trees, time, virtual particles, imaginary numbers, entropy and so on. It is thus pointless to discuss whether a physical concept 'exists' or whether it is 'only' an abstraction used as a tool for descriptions of observations. The two possibilities coincide. The point of dispute can only be whether the description provided by a concept is or is not *precise*.

For mathematical concepts, existence has a somewhat different meaning: a mathematical concept is said to exist if it has no built-in contradictions. This is a much weaker requirement than physical existence. It is thus incorrect to deduce physical existence from mathematical existence. This is a frequent error; from Pythagoras' times onwards it was often stated that since mathematical concepts exist, they must therefore also exist in nature. Historically, this error occurred in the statements that planet orbits 'must' be circles, that planet shapes 'must' be spheres or that physical space 'must' be Euclidean. Today this is still happening with the statements that space and time 'must' be continuous and that nature 'must' be described by sets. In all these cases, the reasoning is wrong. In fact, the continuous attempts to deduce physical existence from mathematical existence hide that the opposite is correct: a short reflection shows that mathematical existence is a special case of physical existence. Challenge 1112 n

We note that there is also a different type of existence, namely *psychological existence*. A concept can be said to exist psychologically if it describes human internal experience. Thus a concept can exist psychologically even if it does not exist physically. It is easy to find examples from the religions or from systems that describe inner experiences. Also myths, legends and comic strips define concepts that only exist psychologically, not physically. In our walk, whenever we talk about existence, we mean physical existence only. Challenge 1113 n

Do things exist?

> Wer Wissenschaft und Kunst besitzt,
> Hat auch Religion;
> Wer jene beiden nicht besitzt,
> Der habe Religion.*
> Johann Wolfgang von Goethe, *Zahme Xenien, IX*

Using the above definition of existence, the question becomes either trivial or imprecise. It is trivial in the sense that things necessarily exist if they describe observations, since they were defined that way. But perhaps the questioner meant to ask: Does reality exist independently of the observer?

Using the above, this question can be rephrased: 'Do the things we observe exist independently of observation?' After thousands of years of extensive discussion by professional philosophers, logicians, sophists and amateurs the answer is the same: it is 'Yes', because the world did not change after great-grandmother died. The disappearance of observers does not seem to change the universe. These experimental findings can be corroborated by inserting the definition of 'existence' into the question, which then becomes: 'Do the things we observe interact with other aspects of nature when they do not interact with people?' The answer is evident. Recent popular books on quantum mechanics fantasize about the importance of the 'mind' of observers – whatever this term may mean; they provide pretty examples of authors who see themselves as irreplaceable, seemingly having lost the ability to see themselves as part of a larger entity.

Of course there are other opinions about the existence of things. The most famous is that of the Irishman George Berkeley (1685–1753) who rightly understood that thoughts based on observation alone, if spread, would undermine the basis of the religious organization of which he was one of the top managers. To counteract this tendency, in 1710 he published *A Treatise Concerning the Principles of Human Knowledge*, a book denying the existence of the material world. This reactionary book became widely known in like-minded circles (it was a time when few books were written) even though it is based on a fundamentally flawed idea: it assumes that the concept of 'existence' and that of 'world' can be defined independently. (You may be curious to try the feat.) Challenge 1114 e

Berkeley had two aims when he wrote his book. First, he tried to deny the capacity of people to arrive at judgements on nature or on any other matter *from their own experience*. Second, he also tried to deny the *ontological reach* of science, i.e. the conclusions one can draw from experience on the questions about human existence. Even though Berkeley is generally despised nowadays, he actually achieved his main aim: he was the originator of the statement that science and religion do not contradict, but *complement* each other. By religion, Berkeley did not mean either morality or spirituality; every scientists is a friend of both of these. By religion, Berkeley meant that the standard set of beliefs that he stood for is above the deductions of reason. This widely cited statement, itself a belief, is still held dearly by many even to this day. However, when searching for the origin of motion, all beliefs stand in the way, including this one. Carrying beliefs is like carrying oversized baggage: it prevents one from reaching the top of Motion Mountain.

* He who possesses science and art, also has religion; he who does not possess the two, better have religion.

Does the void exist?

Teacher: 'What is found between the nucleus and the electrons?'
Student: 'Nothing, only air.'

Natura abhorret vacuum.

Antiquity

In philosophical discussions '*void*' is usually defined as 'non-existence'. It then becomes a game of words to ask for a yes or no answer to the question 'Does the void exist?' The expression 'the existence of non-existence' is either a contradiction of terms or is at least unclearly defined; the topic would not seem to be of great interest. However, similar questions do appear in physics, and a physicist should be prepared to notice the difference of this from the previous one. Does a vacuum exist? Does empty space exist? Or is the world 'full' everywhere, as the more conservative biologist Aristotle maintained? In the past, people have been killed for giving an answer that was unacceptable to authorities.

It is not obvious, but it is nevertheless important, that the modern physical concepts of 'vacuum' and 'empty space' are not the same as the philosophical concept of 'void'. 'Vacuum' is not defined as 'non-existence'; on the contrary, it is defined as the absence of matter and radiation. Vacuum is an entity with specific observable properties, such as its number of dimensions, its electromagnetic constants, its curvature, its vanishing mass, its interaction with matter through curvature and through its influence on decay, etc. (A table of the properties of a physical vacuum is given on page 538.) Historically, it took a long time to clarify the distinction between a physical vacuum and a philosophical void. People confused the two concepts and debated the existence of the vacuum for more than two thousand years. The first to state that it existed, with the courage to try to look through the logical contradiction at the underlying physical reality, were Leucippus and Democritus, the most daring thinkers of antiquity. Their speculations in turn elicited the reactionary response of Aristotle, who rejected the concept of vacuum. Aristotle and his disciples propagated the belief about nature's *horror of the vacuum*.

The discussion changed completely in the seventeenth century, when the first experimental method to realize a vacuum was devised by Torricelli.* Using mercury in a glass tube, he produced the first laboratory vacuum. Can you guess how? Challenge 1115 n Arguments against the existence of the vacuum again appeared around 1900, when it was argued that light needed 'aether' for its propagation, using almost the same arguments that had been used two hundred years earlier, but in different words. However, experiments failed to detect any of the supposed properties of this unclearly defined concept. Experiments in the field of general relativity showed that a vacuum can move – though in a completely different way from the way in which the aether was expected to move – that the vacuum can be bent, but it then tends to return to its shape. Then, in the late twentieth century, quantum field theory again argued against the existence of a true vacuum and in favour of a space full of virtual particle–antiparticle pairs, culminating in the discussions around the cosmological constant. Page 918

* Evangelista Torricelli (b. 1608 Faenza, d. 1647 Florence), Italian physicist, pupil and successor to Galileo. The (non-SI) pressure unit 'torr' is named after him.

The question 'Does the void exist?' is settled conclusively only in the third part of this Page 941 walk, in a rather surprising way.

Is nature infinite?

> It is certain and evident to our senses, that in the world some things are in motion. Now whatever is moved is moved by another... If that by which it is moved be itself moved, then this also needs to be to be moved by another, and that by another again. But this cannot go on to infinity, because then there would be no first mover and consequently, no other mover, seeing that subsequent movers move only inasmuch as they are moved by the first mover, as the staff moves only because it is moved by the hand. Therefore it is necessary to arrive at a first mover, moved by no other; and this everyone understands to be god.
>
> Thomas Aquinas (*c.* 1225–1274) *Summa Theologiae*, I, q. 2.

Most of the modern discussions about set theory centre on ways to defining the term 'set' for various types of infinite collections. For the description of motion this leads to two questions: Is the universe infinite? Is it a set? We begin with the first one. Illuminating the question from various viewpoints, we will quickly discover that it is both simple and imprecise.

Do we need infinite quantities to describe nature? Certainly, in classical and quantum physics we do, e.g. in the case of space-time. Is this necessary? We can say already a few things.

Any set can be finite in one aspect and infinite in another. For example, it is possible to proceed along a finite mathematical distance in an infinite amount of time. It is also possible to travel along any distance whatsoever in a given amount of mathematical time, making infinite speed an option, even if relativity is taken into account, as was explained Page 269 earlier.

Despite the use of infinities, scientists are still limited. We saw above that many types of Page 601 infinities exist. However, no infinity larger than the cardinality of the real numbers plays a role in physics. No space of functions or phase space in classical physics and no Hilbert Ref. 612 space in quantum theory has higher cardinality. Despite the ability of mathematicians to define much larger kinds of infinities, the description of nature does not need them. Even the most elaborate descriptions of motion use only the infinity of the real numbers.

But is it possible at all to say *of nature* or of one of its aspects that it is indeed *infin-* Challenge 1116 n *ite*? Can such a statement be compatible with observations? No. It is evident that every statement that claims that something in nature is infinite is a belief, and is not backed by observations. We shall patiently eliminate this belief in the following.

The possibility of introducing false infinities make any discussion on whether human- Ref. 626 ity is near the 'end of science' rather difficult. The amount of knowledge and the time required to discover it are unrelated. Depending on the speed with which one advances through it, the end of science can be near or unreachable. In practice, scientists have thus the power to *make* science infinite or not, e.g. by reducing the speed of progress. As scientists need funding for their work, one can guess the stand that they usually take.

In short, the universe cannot be proven to be infinite. But can it be *finite*? At first sight, this would be the only possibility left. (It is not, as we shall see.) But even though many have tried to describe the universe as finite in all its aspects, no one has yet been successful. In order to understand the problems that they encountered, we continue with the other question mentioned above:

Is the universe a set?

A simple observation leads us to question whether the universe is a set. For 2500 years Ref. 627 it has been said that the universe is made of vacuum and particles. This implies that the universe is made of a certain *number* of particles. Perhaps the only person to have taken this conclusion to the limit was the English astrophysicist Arthur Eddington (1882–1944), who wrote:

> I believe there are 15,747,724,136,275,002,577,605,653,961,181,555,468,044,717,914-,527,116,709,366,231,425,076,185,631,031,296 protons in the universe and the same number of electrons. Ref. 628

Eddington was ridiculed over and over again for this statement and for his beliefs that lead up to it. His arguments were indeed based on his personal preferences for certain pet numbers. However, we should not laugh too loudly. In fact, for 2500 years almost all scientists have thought along the same line, the only difference being that they have left the precise number unspecified! In fact, *any other number* put into the above sentence would be equally ridiculous. Avoiding specifying it is just a cowards' way of avoiding looking at this foggy aspect of the particle description of nature.

Is there a particle number at all? If you smiled at Eddington's statement, or if you shook your head over it, it may mean that you instinctively believe that nature is not a set. Is this so? Whenever we define the universe as the totality of events, or as the totality of all space-time points and objects, we imply that space-time points can be distinguished, that objects can be distinguished and that both can be distinguished from each other. We always assume that nature is separable and a set. But is this correct? The question is important. The ability to distinguish space-time points and particles from each other is often called *locality*. Thus the universe is separable or a set if and only if our description of it is local.* And in everyday life, locality is observed without exception.

In daily life we also observe that nature is separable and a whole at the same time. It is a 'many that can be thought as one': in daily life nature is a set. Indeed, the basic characteristic of nature is its diversity. In the world around us we observe changes and differences; we observe that nature is separable. Furthermore, all aspects of nature belong together: there are relations between these aspects, often called 'laws,' stating that the different aspects of nature form a whole, usually called the universe.

* In quantum mechanics also other, less clear definitions of locality are used. We will mention them in the second part of this text. The issue mentioned here is a different, more fundamental one, and not connected with that of quantum theory.

In other words, the possibility of describing observations with the help of 'laws' follows from our experience of the separability of nature. The more precisely the separability is specified, the more precisely the 'laws' can be formulated. Indeed, if nature were not separable or were not a unity, we could not explain why stones fall downwards. Thus we are led to speculate that we should be able to deduce all 'laws' from the fact that nature is separable.

In addition, only the separability allows us to describe nature at all. A description is a classification, that is, a mapping between certain aspects of nature and certain concepts. All concepts are sets and relations. Since the universe is separable, it can be described with the help of sets and relations. Both are separable entities with distinguishable parts. A precise description is commonly called an understanding. In short, the universe is comprehensible only because it is separable.

Moreover, only the separability of the universe makes our brain such a good instrument. The brain is built from a large number of connected components, and only the brain's separability allows it to function. In other words, thinking is only possible because nature is separable.

Finally, only the separability of the universe allows us to distinguish reference frames, and thus to define all symmetries at the basis of physical descriptions. And in the same way that separability is thus necessary for *covariant* descriptions, the unity of nature is necessary for *invariant* descriptions. In other words, the so-called 'laws' of nature are based on the experience that nature is both separable and unifiable – that it is a set.

These arguments seem overwhelmingly to prove that the universe is a set. However, these arguments apply only to everyday experience, everyday dimensions and everyday energies. Is nature a set also *outside* the domains of daily life? Are objects different at all energies, even when they are looked at with the highest precision possible? We have three open issues left: the issue of the number of particles in the universe; the circular definition of space, time and matter; and the issue as to whether describing nature as made of particles and void is an overdescription, an underdescription, or neither. These three issues make us doubt whether objects are countable at all energies. We will discover in the
Page 966 third part of our mountain ascent that this is not the case in nature. The consequences will be extensive and fascinating. As an example, try to answer the following: if the universe
Challenge 1117 n is not a set, what does that mean for space and time?

Does the universe exist?

> Each progressive spirit is opposed by a thousand men appointed to guard the past.
> Maurice Maeterlink (1862–1949), Belgian dramatist

Following the definition above, existence of a concept means its usefulness to describe interactions. There are two common definitions of the concept of 'universe'. The first is the totality of all matter, energy and space-time. But this usage results in a strange consequence: since nothing can interact with this totality, we cannot claim that the universe exists.

So let us take the more restricted view, namely that the universe is only the totality of all matter and energy. But also in this case it is impossible to interact with the universe.

Challenge 1118 n Can you give a few arguments to support this?

In short, we arrive at the conclusion that the universe does not exist. We will indeed Page 969 confirm this result in more detail later on in our walk. In particular, since the universe does not exist, it does not make sense to even try to answer *why* it exists. The best answer Ref. 586 might be: because of furiously sleeping, colourless green ideas.

What is creation?

> (Gigni) De nihilo nihilum, in nihilum nil posse reverti.*
>
> Persius, *Satira*, III, v. 83-84. Ref. 629

> Anaxagoras, discovering the ancient theory that nothing comes from nothing, decided to abolish the concept of creation and introduced in its place that of discrimination; he did not hesitate to state, in effect, that all things are mixed to the others and that discrimination produces their growth.
>
> *Anonymous fragment*, middle ages. Ref. 630

The term 'creation' is often heard when talking about nature. It is used in various contexts with different meanings.

One speaks of creation as the characterization of human actions, such as observed in an artist painting or a secretary typing. Obviously, this is one type of change. In the classification of change introduced at the beginning of our walk, the changes cited are movements of objects, such as the electrons in the brain, the molecules in the muscles, the material of the paint, or the electrons inside the computer. This type of creation is thus a special case of motion.

One also speaks of creation in the biological or social sense, such as in 'the creation of life', or 'creation of a business', or 'the creation of civilization'. These events are forms of growth or of self-organization; again, they are special cases of motion.

Physicists one often say that a lamp 'creates' light or that a stone falling into a pond 'creates' water ripples. Similarly, they talk of 'pair creation' of matter and antimatter. It Page 708 was one of the important discoveries of physics that all these processes are special types Page 792 of motion, namely excitation of fields.

In popular writing on cosmology, 'creation' is also a term commonly applied, or better misapplied, to the *big bang*. However, the expansion of the universe is a pure example of motion, and contrary to a frequent misunderstanding, the description of the big bang contains no process that does not fall into one of the previous three categories, as shown in Page 429 the chapter on general relativity. Quantum cosmology provides more reasons to support Page 428 the fact that the term 'creation' is not applicable to the big bang. First, it turns out that the big bang was not an event. Second, it was not a beginning. Third, it did not provide a *choice* from a large set of possibilities. The big bang does not have any properties attributed to the term 'creation'.

* Nothing (can appear) from nothing, nothing can disappear into nothing.

In summary, we conclude that in all cases, *creation is a type of motion*. (The same applies to the notions of 'disappearance' and 'annihilation'.) No other type of creation is observed in nature. In particular, the naive sense of 'creation', namely 'appearance from nothing' – *ex nihilo* in Latin – is never observed in nature. All observed types of 'creation' require space, time, forces, energy and matter for their realization. Creation requires something to exist already, in order to take place. In addition, precise exploration shows that no physical process and no example of motion has a beginning. Our walk will show us that nature does not allow us to pinpoint beginnings. This property alone is sufficient to show that 'creation' is not a concept applicable to what happens in nature. Worse still, creation is applied only to physical systems; we will discover that nature is not a system and that systems do not exist.

The opposite of creation is *conservation*. The central statements of physics are conservation theorems: for energy, mass, linear momentum, angular momentum, charge, etc. In fact, every conservation 'law' is a detailed and accurate rejection of the concept of creation. The ancient Greek idea of atoms already contains this rejection. Atomists stated that there is no creation and no disappearance, but only motion of atoms. Every transformation of matter is a motion of atoms. In other words, the idea of the atom was a direct consequence of the negation of creation. It took humanity over 2000 years before Page 203 it stopped locking people in jail for talking about atoms, as had happened to Galileo.

However, there is one exception in which the naive concept of creation does apply: it describes what magicians do on stage. When a magician makes a rabbit appear from nowhere, we indeed experience 'creation' from nothing. At its best such magic is a form of entertainment, at its worst, a misuse of gullibility. The idea that the universe results from either of these two does not seem appealing; on second thought though, maybe looking at the universe as the ultimate entertainment could open up a fresh and more productive approach to life.

Voltaire (1694–1778) popularized an argument against creation often used in the past: we do not know whether creation has taken place or not. Today the situation is different: we *do* know that it has *not* taken place, because creation is a type of motion and, as we will see in the third part of our mountain ascent, motion did not exist near the big bang.

Have you ever heard the expression 'creation of the laws of nature'? It is one of the most common examples of disinformation. First of all, this expression confuses the 'laws' with nature itself. A description is not the same as the thing itself; everybody knows that giving their beloved a description of a rose is different from giving an actual rose. Second, the expression implies that nature is the way it is because it is somehow 'forced' to follow the 'laws' – a rather childish and, what is more, incorrect view. And third, the expression assumes that it is possible to 'create' descriptions of nature. But a 'law' is a description, and a description by definition cannot be created: so the expression makes no sense at all. The expression 'creation of the laws of nature' is the epitome of confused thinking.

It may well be that calling a great artist 'creative' or 'divine', as was common during the Renaissance, is not blasphemy, but simply an encouragement to the gods to try to do as well. In fact, whenever one uses the term 'creation' to mean anything other than some form of motion, one is discarding both observations and human reason. It is one of the last pseudo-concepts of our modern time; no expert on motion should forget this. It is impossible to escalate Motion Mountain without getting rid of 'creation'. This is not Page 890 easy. We will encounter the next attempt to bring back creation in the study of quantum theory.

> Every act of creation is first of all an act of destruction.
>
> Pablo Picasso (1881–1973), painter.

Is nature designed?

> In the beginning the universe was created. This has made a lot of people very angry and has been widely regarded as a bad move.
>
> Douglas Adams

The tendency to infer the creation of an object from its simple existence is widespread. Some people jump to this conclusion every time they see a beautiful landscape. This habit stems from the triple prejudice that a beautiful scene implies a complex description, in turn implying complex building instructions, and therefore pointing to an underlying *design*.

This chain of thought contains several mistakes. First, beauty is not necessarily a consequence of complexity. Usually it is the opposite; indeed, the study of chaos and of self-organization demonstrated how beautifully complex shapes and patterns can be generated with extremely simple descriptions. True, for most human artefacts, complex descriptions indeed imply complex building processes; a personal computer is a good example of a complex object with a complex production process. But in nature, this connection does not apply. We have seen above that even the amount of information needed to construct a human body is about a million times smaller than the information stored in the brain alone. Similar results have been found for plant architecture and for many other examples of patterns in nature. The simple descriptions behind the apparent complexities of nature have been and are still being uncovered by the study of self-organization, chaos, turbulence and fractal shapes. In nature, complex structures derive from *simple* processes. Beware of anyone who says that nature has 'infinite complexity': first of all, complexity is not a measurable entity, despite many attempts to quantify it. In addition, all known complex system can be described by (relatively) few parameters and simple equations. Finally, nothing in nature is infinite.

Ref. 631

The second mistake in the argument for design is to link a description with an 'instruction', and maybe even to imagine that some unknown intelligence is somehow pulling the strings of the world's stage. The study of nature has consistently shown that there is no hidden intelligence and no instruction behind the processes of nature. An instruction is a list of orders to an executioner. But there are no orders in nature, and no executioners. There are no 'laws' of nature, only descriptions of processes. Nobody is building a tree; the tree is an outcome of the motion of molecules making it up. The genes in the tree do contain information; but no molecule is given any instructions. What seem to be instructions to us are just natural movements of molecules and energy, described by the same patterns taking place in non-living systems. The whole idea of instruction – like that of 'law' of nature – is an ideology, born from an analogy with monarchy or even tyranny, and a typical anthropomorphism.

The third mistake in the argument for design is the suggestion that a complex description for a system implies an underlying design. This is not correct. A complex description only implies that the system has a long evolution behind it. The correct deduction is: something of large complexity exists; therefore it has *grown*, i.e. it has been transformed through input of (moderate) energy over time. This deduction applies to flowers, mountains, stars, life, people, watches, books, personal computers and works of art; in fact it applies to all objects in the universe. The complexity of our environment thus points out the considerable age of our environment and the shortness of our own life.

In summary, the lack of basic complexity and the lack of instructions in nature confirm a simple result: there is not a single observation in nature that implies or requires design or creation. On the other hand, the variety and intensity of nature's phenomena fills us with deep awe. The wild beauty of nature shows us how *small* a part of nature we actually are, both in space and in time.* We shall explore this experience in detail. We shall find that remaining open to nature's phenomena in all their overwhelming intensity is central to the rest of our adventure.

Page 890

> There is a separation between state and church, but not yet between state and science.
>
> Paul Feyerabend (1924–1994)

What is a description?

> In theory, there is no difference between theory and practice. In practice, there is.

Following standard vocabulary usage, a description of an observation is a list of the details. The above example of the grampus showed this clearly. In other words, a *description* of an observation is the act of categorizing it, i.e. of comparing, by identifying or distinguishing, the observation with all the other observations already made. A description is a classification. In short, *to describe means to see as an element of a larger set*.

Page 628

A description can be compared to the 'you are here' sign on a city tourist map. Out of a set of possible positions, the 'you are here' sign gives the actual one. Similarly, a description highlights the given situation in comparison with all other possibilities. For example, the formula $a = GM/r^2$ is a description of the observations relating motion to gravity, because it classifies the observed accelerations a according to distance to the central body r and to its mass M; indeed such a description sees each specific case as an example of a general pattern. The habit of generalizing is one reason for the often disturbing dismissiveness of scientists: when they observe something, their professional training usually makes them classify it as a special case of a known phenomenon and thus keeps them from being surprised or from being exited about it.

* The search for a 'sense' in life or in nature is a complicated (and necessary) way to try to face the smallness of human existence.

A description is thus the opposite of a *metaphor*; the latter is an analogy relating an observation with another *special* case; a description relates an observation with a *general* case, such as a physical theory.

> Felix qui potuit rerum cognoscere causas,
> atque metus omnis et inexorabile fatum
> subjecit pedibus strepitumque acherontis avari.
>
> Vergilius*

Reason, purpose and explanation

> Der ganzen modernen Weltanschauung liegt die Täuschung zugrunde, daß die sogenannten Naturgesetze die Erklärungen der Naturerscheinungen seien.*
>
> Ludwig Wittgenstein, *Tractatus*, 6.371

▪ Why are the leaves of most trees green? Because they absorb red and blue light. Why do they absorb those colours? Because they contain chlorophyll. Why is chlorophyll green? Because all chlorophyll types contain magnesium between four pyrrole groups, and this chemical combination gives the green colour, as a result of its quantum mechanical energy levels. Why do plants contain chlorophyll? Because this is what land plants can synthesize. Why only this? Because all land plants originally evolved from the green algae, who are only able to synthesize this compound, and not the compounds found in the blue or in the red algae, which are also found in the sea.

▪ Why do children climb trees, and why do some people climb mountains? Because of the sensations they experience during their activity: the feelings of achievement, the symbolic act to go upwards, the wish to get a wider view of the world are part of this type of adventure.

The 'why'-questions in the last two paragraphs show the general difference between reasons and purposes (although the details of these two terms are not defined in the same way by everybody). A *purpose* or *intention* is a classification applied to the actions of humans or animals; strictly speaking, it specifies the quest for a feeling, namely for achieving some type of satisfaction after completion of the action. On the other hand, a *reason* is a specific relation of a fact with the rest of the universe, usually its past. What we call a reason always rests outside the observation itself, whereas a purpose is always internal to it.

Reasons and purposes are the two possibilities of explanations, i.e. the two possible answers to questions starting with 'why'. Usually, physics is not concerned with purpose or with people's feeling, mainly because its original aim, to talk about motion with precision,

* 'Happy he who can know the causes of things and who, free of all fears, can lay the inexorable fate and the noise of Acheron to his feet.' (Georg. 2, 490 ss.) Publius Vergilius Maro (70–19 BCE), the great roman poet, is author of the Aeneis. Acheron was the river crossed by those who had just died and were on their way to the Hades.

* The whole modern conception of the world is founded on the illusion that the so-called laws of nature are the explanations of natural phenomena.

does not seem to be achievable in this domain. Therefore, *physical* explanations of facts are never purposes, but are always reasons. A *physical explanation* of an observation is always the description of its relation with the rest of nature.* Ref. 632

This means that – contrary to common opinion – a question starting with 'why' is accessible to physical investigation as long as it asks for a reason and not for a purpose. In particular, questions such as 'why do stones fall downwards and not upwards?' or 'why do electrons have that value of mass, and why do they have mass at all?' or 'why does space have three dimensions and not thirty-six?' can be answered, as these ask for the connection between specific observations and more general ones. Of course, not all demands for explanation have been answered yet, and there are still problems to be solved. Our present trail only leads from a few answers to some of the more fundamental questions about motion.

The most general quest for an explanation derives from the question: why is the universe the way it is? The topic is covered in our mountain ascent using the two usual approaches, namely:

Unification and demarcation

> Tout sujet est un; et, quelque vaste qu'il soit, il peut être renfermé dans un seul discours.*
>
> Buffon (1701–1780), *Discours sur le style*.

Studying the properties of motion, constantly paying attention to increase the accuracy of description, we find that explanations are generally of two types:**

1 ▪ 'It is like all such cases; also this one is described by ...' The situation is recognized as a *special case* of a general behaviour.

1 ▪ 'If the situation were different, we would have a conclusion in contrast with observations.' The situation is recognized as the *only possible case*.***

In other words, the first approach is to formulate rules or 'laws' that describe larger and larger numbers of observations, and compare the observation with them. This endeavour is called the *unification* of physics – by those who like it; those who don't like it, call it 'reductionism'. For example, the same rule describes the flight of a tennis ball, the motion of the tides at the sea shore, the timing of ice ages, and the time at which the planet Venus ceases to be the evening star and starts to be the morning star. These processes are all consequences of universal gravitation. Similarly, it is not evident that the same rule describes the origin of the colour of the eyes, the formation of lightning, the digestion of food and the working of the brain. These processes are described by quantum electrodynamics.

---

* It is important to note that purposes are *not* put aside because they pertain to the future, but because they are inadmissible anthropomorphisms. In fact, for deterministic systems, we can equally say that the future is actually a *reason* for the present and the past, a fact often forgotten.

* Every subject is one and, however vast it is, it can be comprised in a single discourse.

Challenge 1119 n ** Are these the only possible ones?

*** These two cases have not to be confused with similar sentences that *seem* to be explanations, but that aren't:

▪ 'It is like the case of ...' A similarity with another *single* case is *not* an explanation.

▪ 'If it were different, it would contradict the idea that ...' A contradiction with an *idea* or with a theory is *not* an explanation.

Unification has its most impressive successes when it predicts an observation that has not been made before. A famous example is the existence of antimatter, predicted by Dirac when he investigated the solutions of an equation that describes the precise behaviour of common matter.

The second procedure in the search for explanations is the elimination of all other imaginable alternatives in favour of the actually correct one. This endeavour has no commonly accepted name: it could be called the *demarcation* of the 'laws' of physics – by those who like it; others call it 'anthropocentrism', or simply 'arrogance'.

When we discover that light travels in such a way that it takes the shortest possible time to its destination, when we describe motion by a principle of least action, or when we discover that trees are branched in such a way that they achieve the largest effect with the smallest effort, we are using a demarcation viewpoint.

In summary, unification, answering 'why' questions, and demarcation, answering 'why not' questions, are typical for the progress throughout the history of physics. We can say that the dual aspects of unification and demarcation form the composing and the opposing traits of physics. They stand for the desire to *know everything*.

However, neither demarcation nor unification can explain the universe. Can you see why? In fact, apart from unification and demarcation, there is a third possibility that merges the two and allows one to say more about the universe. Can you find it? Our walk will automatically lead to it later.

Challenge 1120 n
Challenge 1121 n

Pigs, apes and the anthropic principle

> Das wichtigste Instrument des Wissenschaftlers ist der Papierkorb.*

The wish to achieve demarcation of the patterns of nature is most interesting when we follow the consequences of different rules of nature until we find them in contradiction with the most striking observation: our own human existence. In this special case the program of demarcation is often called the *anthropic principle* – from the Greek ἄνθρωπος, meaning 'man'.

For example, if the Sun–Earth distance were different from what it is, the resulting temperature change on the Earth would have made impossible the appearance of life, which needs liquid water. Similarly, our brain would not work if the Moon did not circle the Earth. It is only because the Moon revolves around our planet that the Earth's magnetic field is large enough to protect the Earth by deviating most of the cosmic radiation that would otherwise make all life on Earth impossible. It is only because the Moon revolves around our planet that the Earth's magnetic field is small enough to leave enough radiation to induce the mutations necessary for evolution. It is also well-known that if there were fewer large planets in the solar system, the evolution of humans would have been impossible. The large planets divert large numbers of comets, preventing them from hitting the Earth. The spectacular collision of comet Shoemaker–Levy-9 with Jupiter, the astronomical event of July 1994, was an example of this diversion of a comet.**

* The most important instrument of a scientist is the waste paper basket.
** For a collection of pictures of this event, see e.g. the http://garbo.uwasa.fi/pc/gifslevy.html website.

Also the anthropic principle has its most impressive successes when it predicts unknown observations. The most famous example stems from the study of stars. Carbon atoms, like all other atoms except most hydrogen, helium or lithium atoms, are formed in stars through fusion. While studying the mechanisms of fusion in 1953, the well-known British astrophysicist Fred Hoyle* found that carbon nuclei could not be formed from the alpha particles present inside stars at reasonable temperatures, unless they had an excited state with an increased cross-section. From the fact of our existence, which is based on carbon, Hoyle thus predicted the existence of a previously unknown excited state of the carbon nucleus. And, indeed, the excited state was found a few months later by Willy Fowler.** Ref. 633

In its *serious* form, the anthropic principle is therefore the quest to deduce the description of nature from the experimental fact of our own existence. Ref. 634 In the popular literature, however, the anthropic principle is often changed from a simple experimental method to deduce the patterns of nature, to its *perverted* form, a melting pot of absurd metaphysical ideas in which everybody mixes up their favourite beliefs. Most frequently, the experimental observation of our own existence has been perverted to reintroduce the idea of 'design', i.e. that the universe has been constructed with the aim of producing humans; often it is even suggested that the anthropic principle is an *explanation* – a gross example of disinformation.

How can we distinguish between the serious and the perverted form? We start with an observation. We would get exactly the same rules and patterns of nature if we used the existence of pigs or monkeys as a starting point. In other words, if we would reach *different* conclusions by using the *porcine principle* or the *simian principle*, we are using the perverted form of the anthropic principle, otherwise we are using using the serious form. (The carbon-12 story is thus an example of the serious form.) This test is effective because there is no known pattern or 'law' of nature that is particular to humans but unnecessary apes or for pigs.*

> Er wunderte sich, daß den Katzen genau an den Stellen Löcher in den Pelz geschnitten wären, wo sie Augen hätten.
>
> Georg Christoph Lichtenberg**

* Fred Hoyle (b. 1915 Bingley, Yorkshire, d. 2001), important British astronomer and astrophysicist. He was the first and maybe only physicist who ever made a specific prediction – namely the existence of an excited state of the carbon nucleus – from the simple fact that humans exist. A permanent maverick, he coined the term 'big bang' even though he did not accept the evidence for it, and proposed another model, the 'steady state'. His most important and well-known research was on the formation of atoms inside stars. He also propagated the belief that life was brought to Earth from extraterrestrial microbes.

** William A. Fowler (1911–1995) shared the 1983 Nobel prize in physics with Subramanyan Chandrasekhar for this and related discoveries.

* Though apes do not seem to be good physicists, as described in the text by D.J. Povinelli, *Folk Physics for Apes: the Chimpanzee's Theory of How the World Works*, Oxford University Press, 2000.

** 'He was amazed that cats had holes cut into their fur precisely in those places where they had eyes.' Georg Christoph Lichtenberg (1742–1799), German physicist and intellectual, professor in Göttingen, still famous today for his extremely numerous and witty aphorisms and satires. Among others of his time, Lichtenberg made fun of all those who maintained that the universe was made exactly to the measure of man, a frequently encountered idea in the foggy world of the anthropic principle.

Does one need cause and effect in explanations?

> In nature there are neither rewards nor punishments – there are consequences.
>
> Ivan Illich (b. 1926 Vienna, d. 2002 Bremen)

> The world owes you nothing. It was there first.
>
> Mark Twain (1835–1910)

> No matter how cruel and nasty and evil you may be, every time you take a breath you make a flower happy.
>
> Mort Sahl

Historically, the two terms 'cause' and 'effect' have played an important role in philosophical discussions. In particular, during the birth of modern mechanics, it was important to point out that every effect has a cause, in order to distinguish precise thought from thought based on beliefs, such as 'miracles', 'divine surprises' or 'evolution from nothing'. It was equally essential to stress that effects are different from causes; this distinction avoids pseudo-explanations such as the famous example by Molière where the doctor explains to his patient in elaborate terms that sleeping pills work because they contain a dormitive virtue.

Ref. 635

But in physics, the concepts of cause and effect are not used at all. That miracles do not appear is expressed every time we use symmetries and conservation theorems. The observation that cause and effect differ from each other is inherent in any evolution equation. Moreover, the concepts of cause and effect are not clearly defined; for example, it is especially difficult to define what is meant by one cause as opposed to several of them, and the same for one or several effects. Both terms are impossible to quantify and to measure. In other words, useful as 'cause' and 'effect' may be in personal life for distinction between events that regularly succeed each other, they are not necessary in physics. In physical explanations, they play no special roles.

> Ἀγαθον καὶ ξαξόν · ἓν καὶ ταὐτό.*
>
> Heraclitus

> Wenn ein Arzt hinter den Sarg seines Patienten geht, so folgt manchmal tatsächlich die Ursache der Wirkung.**
>
> Robert Koch, (1843–1910) medical researcher.

Is consciousness required?

Variatio delectat.*

Cicero

Ref. 636 A lot of mediocre discussions are going on about this topic, and we will skip them here. What is consciousness? Most simply and concretely, consciousness means the possession of a small part of oneself that is watching what the rest of oneself is perceiving, feeling, thinking and doing. In short, consciousness is the ability to observe oneself, and in particular one's inner mechanisms and motivations. *Consciousness* is the ability of introspection. For this reason, consciousness is *not* a prerequisite for studying motion. Indeed, animals, plants or machines are also able to observe motion. For the same reason, consciousness is not necessary to observe quantum mechanical motion. On the other hand, both the study of motion and that of oneself have a lot in common: the need to observe carefully, to overcome preconceptions, to overcome fear and the fun of doing so.

For the time being, we have put enough emphasis on the precision of concepts. Talking about motion is also something to be deeply enjoyed. Let us see why.

Precision and clarity obey the indeterminacy relation: their product is constant.

Curiosity

Precision is the child of curiosity.

Like the history of every person, also the history of mankind charts a long struggle to avoid the pitfalls of accepting the statements of authorities as truth, without checking the facts. Indeed, whenever curiosity leads us to formulate a question, there are always two general ways to proceed. One is to check the facts personally, the other is to ask somebody. However, the last way is dangerous: it means to give up a part of oneself. Healthy people, children whose curiosity is still alive, as well as scientists, choose the first way. After all, science is adult curiosity.

Curiosity, also called the *exploratory drive*, plays strange games with people. Starting with the original experience of the world as a big 'soup' of interacting parts, curiosity can drive one to find *all* the parts and *all* the interactions. It drives not only people. It has been observed that when rats show curious behaviour, certain brain cells in the hypothalamus get active and secrete hormones that produce positive feelings and emotions. If a rat has the possibility, via some implanted electrodes, to excite these same cells by pressing a switch, it does so voluntarily: rats get *addicted* to the feelings connected with curiosity. Ref. 637 Like rats, humans are curious because they enjoy it. They do so in at least four ways:

* 'Good and bad – one and the same.'
** 'When a doctor walks behind the coffin of his patient, indeed the cause sometimes follows the effect.'
* 'Change pleases.' Marcus Tullius Cicero (106–43 BCE), important lawyer, orator and politician at the end of the Roman republic.

because they are artists, because they are fond of pleasure, because they are adventurers and because they are dreamers. Let us see how.

Originally, curiosity stems from the desire to interact in a positive way with the environment. Young children provide good examples: curiosity is a natural ingredient of their life, in the same way that it is for other mammals and a few bird species; incidentally, the same taxonomic distribution is found for play behaviour. In short, all animals that play are curious, and vice versa. Curiosity provides the basis for learning, for creativity and thus for every human activity that leaves a legacy, such as art or science. The artist and art theoretician Joseph Beuys (1920–1986) had as his own guiding principle that *every* creative act is a form of art. Humans, and especially children, enjoy curiosity because they feel its importance for creativity, and for growth in general.

Ref. 638

Curiosity regularly leads one to exclaim: 'Oh!', an experience that leads to the second reason to be curious: relishing feelings of wonder and surprise. Epicurus (Epikuros) (341–271 BCE) maintained that this experience, θαυμάζειν, is the origin of philosophy. These feelings, which nowadays are variously called religious, spiritual, numinous, etc., are the same as those to which rats can become addicted. Among these feelings, Rudolf Otto has introduced the now classical distinction into the fascinating and the frightening. He named the corresponding experiences 'mysterium fascinans' and 'mysterium tremendum'.* Within these distinctions, physicists, scientists, children and connoisseurs take a clear stand: they choose the fascinans as the starting point for their actions and for their approach to the world. Such feelings of fascination induce some children who look at the night sky to dream about becoming astronomers, some who look through a microscope to become biologists or physicists, and so on. (It could also be that genetics plays a role in this pleasure of novelty seeking.)

Ref. 639

Perhaps the most beautiful moments in the study of physics are those appearing after new observations have shaken our previously held thinking habits, have forced us to give up a previously held conviction, and have engendered the feeling of being lost. When, in this moment of crisis, we finally discover a more adequate and precise description of the observations, which provide a better insight into the world, we are struck by a feeling usually called illumination. Anyone who has kept alive the memory and the taste for these magic moments knows that in these situations one is pervaded by a feeling of union between oneself and the world.** The pleasure of these moments, the adventures of the change of thought structures connected with them, and the joy of insight following them provides the drive for many scientists. Little talk and lots of pleasure is their common denominator. In this spirit, the well-known Austrian physicist Victor Weisskopf (1908–2002) liked to say jokingly: 'There are two things that make life worth living: Mozart and quantum mechanics.'

* This distinction is the basis of RUDOLF OTTO, *Das Heilige – Über das Irrationale in der Idee des Göttlichen und sein Verhältnis zum Rationalen*, Beck, München, 1991. This is a new edition of the epoch-making work originally published at the beginning of the twentieth century. Rudolf Otto (1869–1937) was one of the most important theologians of his time.

** Several researchers have studied the situations leading to these magic moments in more detail, notably the Prussian physician and physicist Hermann von Helmholtz (1821–1894) and the French mathematician Henri Poincaré (1854–1912). They distinguish four stages in the conception of an idea at the basis of such a magic moment: saturation, incubation, illumination and verification.

Ref. 640

The choice of moving away from the tremendum towards the fascinans stems from an innate desire, most obvious in children, to reduce uncertainty and fear. This drive is the father of all adventures. It has a well-known parallel in ancient Greece, where the first men studying observations, such as Epicurus, stated explicitly that their aim was to free people from unnecessary fear by deepening knowledge and transforming people from frightened passive victims into fascinated, active and responsible beings. Those ancient thinkers started to popularize the idea that, like the common events in our life, the rarer events also follow rules. For example, Epicurus underlined that lightning is a natural phenomenon caused by interactions between clouds, and stressed that it was a natural process, i.e. a process that followed rules, in the same way as the falling of a stone or any other familiar process of everyday life.

Investigating the phenomena around them, philosophers and later scientists succeeded in freeing people from most of their fears caused by uncertainty and a lack of knowledge about nature. This liberation played an important role in the history of human culture and still pervades in the personal history of many scientists. The aim to arrive at stable, rock-bottom truths has inspired (but also hindered) many of them; Albert Einstein is a well-known example for this, discovering relativity, helping to start up but then denying quantum mechanics.

Interestingly, in the experience and in the development of every human being, curiosity, and therefore the sciences, appears *before* magic and superstition. Magic needs deceit to be effective, and superstition needs indoctrination; curiosity doesn't need either. Conflicts of curiosity with superstitions, ideologies, authorities or the rest of society are thus preprogrammed.

Curiosity is the exploration of limits. For every limit, there are two possibilities: the limit can turn out to be real or apparent. If the limit is real, the most productive attitude is that of acceptance. Approaching the limit then gives strength. If the limit is only apparent and in fact non-existent, the most productive attitude is to re-evaluate the mistaken view, extract the positive role it performed, and then cross the limit. Distinguishing between real and apparent limits is only possible when the limit is investigated with great care, openness and unintentionality. Most of all, exploring limits need courage.

Das gelüftete Geheimnis rächt sich.*
Bert Hellinger (1925–)

* The unveiled secret takes revenge.

Courage

It is dangerous to be right in matters on which the established authorities are wrong.
Voltaire (1694–1778)

Manche suchen Sicherheit, wo Mut gefragt ist, und suchen Freiheit, wo das Richtige keine Wahl läßt.*
Bert Hellinger (1925–)

Most of the material in this intermezzo is necessary in the adventure to get to the top of Motion Mountain. But we need more. Like any enterprise, curiosity also requires courage, and complete curiosity, as aimed for in our quest, requires complete courage. In fact, it is easy to get discouraged on this trip. The journey is often dismissed by others as useless, uninteresting, childish, confusing, damaging or, most often, evil. For example, between the death of Socrates in 399 BCE and Paul Thierry, Baron d'Holbach, in the eighteenth century, no book was published with the statement 'gods do not exist', because of the threats to the life of anyone who dared to make the point. Even today, this type of attitude still abounds, as the newspapers show. Ref. 646

Through the constant elimination of uncertainty, both curiosity and scientific activity are implicitly opposed to any idea, person or organization that tries to avoid the comparison of statements with observations. These 'avoiders' demand to live with superstitions and beliefs. But superstitions and beliefs produce unnecessary fear. And fear is the basis of all unjust authorities. One gets into a vicious circle: avoiding comparison with observation produces fear – fear keeps unjust authority in place – unjust authority avoids comparison with observation – etc.

As a consequence, curiosity and science are fundamentally opposed to unjust authority, a connection that made life difficult for people such as Anaxagoras (500–428 BCE) in ancient Greece, Hypatia in the Christian Roman empire, Galileo Galilei in the church state, Antoine Lavoisier in France and Albert Einstein in Germany. In the second half of the twentieth century, victims were Robert Oppenheimer, Melba Phillips and Chandler Davis in the United States and Andrei Sakharov in the Soviet Union. Each of them tell a horrible but instructive story, as have, more recently, Fang Lizhi, Xu Liangying, Liu Gang and Wang Juntao in China, Kim Song-Man in South Korea, Otanazar Aripov in Uzbekistan, Ramadan al-Hadi al-Hush in Libya, Bo Bo Htun in Burma, as well as many hundreds of others. In many authoritarian societies the antagonism between curiosity and injustice has hindered or even completely suppressed the development of physics and other sciences, with extremely negative economic, social and cultural consequences.

When embarking on this ascent, we need to be conscious of what we are doing. In fact, external obstacles can be avoided or at least largely reduced by keeping the project to oneself. Other difficulties still remain, this time of personal nature. Many have tried to embark on this adventure with some hidden or explicit intention, usually of an ideological nature, and then have got entangled by it before reaching the end. Some have not been

* 'Some look for security where courage is required and look for freedom where the right way doesn't leave any choice.' This is from the beautiful booklet by BERT HELLINGER, *Verdichtetes*, Carl-Auer Systeme Verlag, 1996.

prepared to accept the humility required for such an endeavour. Others were not prepared for the openness required, which can shatter deeply held beliefs. Still others were not ready to turn towards the unclear, the dark and the unknown, confronting them at every occasion.

On the other hand, the dangers are worth it. By taking curiosity as a maxim, facing disinformation and fear with all one's courage, one achieves freedom from all beliefs. In exchange, you come to savour the fullest pleasures and the deepest satisfaction that life has to offer.

We thus continue our hike. At this point, the trail towards the top of Motion Mountain is leading us towards the next adventure: discovering the origin of sizes and shapes in nature.

> And the gods said to man: 'Take what you want, and pay the price.'
>
> (Popular saying)

> It is difficult to make a man miserable while he feels he is worthy of himself.
>
> Abraham Lincoln (1809–1865) US President

Bibliography

577 See for example the beautiful textbook by Stephen C. Stearns & Rolf F. Hoekstra, *Evolution: An Introduction*, Oxford University Press, 2000. For fascinating story of evolution for non-specialists, see Richard Fortey, *Life – An Unauthorized Biography*, Harper Collins, 1997. See also Stephen J. Gould, *The Panda's thumb*, W.W. Norton & Co., New York, 1980, one of the several interesting and informative books on evolutionary biology by the best writer in the field. An informative overview over the results of evolution, with the many-branched family tree that it produced, is given on the http://phylogeny.arizona.edu/tree website. About the results of evolution for human beings, see the informative text by K. Kusch & S. Kusch, *Der Mensch in Zahlen*, Spektrum Akademischer Verlag, 2nd edn., 2000. The epochal work by Charles Darwin, *On the Origin of Species*, can be found on the web, e.g. on the http://entisoft.earthlink.net/origspec.htm website. Cited on page 585.

578 A simple description is Malcolm Ross Macdonald, *The Origin of Johnny*, Jonathan Cape, 1976. Cited on page 585.

579 There is disagreement among experts about the precise timing of this experience. Some say that only birth itself is that moment. However, there are several standard methods to recall memories of early life, even of the time before birth; one is described and used by J.K. Stettbacher, *Making Sense of Suffering*, Penguin, New York, 1991, translated from the German original. Even more impressive examples are given by N.J. Mayer, *Der Kainkomplex – neue Wege der systemischen Familientherapie*, Integral Verlag, 1998. Cited on page 585.

580 Sanjida O'Connell, *Mindreading – How We Learn to Love and Lie*, Arrow, 1998. This interesting book describes the importance of lying in the development of a human being, and explains the troubles those people have who cannot read other minds and thus cannot lie, such as autistics. Cited on pages 586 and 615.

581 The approach to describe observations as related parts is called structuralism; the starting point for this movement was de Saussure's *Cours de linguistique générale* (see the footnote on page 594). A number of thinkers have tried to use the same approach in philosophy, mythology and literature theory, though with little success. An overview of the (modest) success of structuralism in linguistics and its failure in other fields is given by L. Jackson, *The Poverty of Structuralism: Literature and Structuralist Theory*, Longman, 1991. The author argues that when one reduces systems to interactions, one neglects the specific content and properties of the elements of the system, and this approach prevents a full understanding of the system under discussion. Cited on page 587.

582 For a view of the mental abilities different from that of Piaget (described on page 587), a presently much discussed author is the Soviet experimental psychologist Lev Vigotsky, whose path-breaking ideas and complicated life are described, e.g., in Lev Vigotsky, *Mind in Society*, Harvard University Press, 1978, or in René van der Veer & Jaan Valsiner, *Understanding Vigotsky: a Quest for Synthesis*, Blackwell Publishers, 1994. More extensive material can be found in the extensive work by René van der Veer & Jaan Valsinger, *The Vigotsky Reader*, Blackwell, 1994. Cited on page 587.

583 A somewhat unconventional source for more details is the beautiful text by Bruno Bettelheim, The Uses of Enchantment: the Meaning and Importance of Fairy Tales, Knopf, 1976. Cited on page 587.

584 Quoted in V. Harlen, R. Rappmann & P. Schata, *Soziale Plastik – Materialien zu Joseph Beuys*, Achberger Verlag, 1984, p. 61. Cited on page 588.

585 The problems appearing when one loses the ability to classify or to memorise are told in the beautiful book by the neurologist Oliver Sacks, *The Man Who Mistook His Wife for a Hat*, Picador, 1985, which collects many case studies he encountered in his work. More astonishing cases are collected in his equally impressive text *An Anthropologist on Mars*, Picador, 1995.

See also the beautiful text Donald D. Hoffman, *Visual Intelligence – How We Create What We See*, W.W. Norton & Co., 1998, and the http://ari.ss.uci.edu/cogsci/personnel/hoffman.html website associated to it. Cited on pages 588 and 590.

586 For a passionate introduction to the connections between language and the brain from a Chomskian perspective, see the bestselling book by Steven Pinker, *The Language Instinct – How the Mind Creates Language*, Harper Perennial, 1994. The green idea sentence is discussed in a chapter of the book. Cited on pages 588, 620, and 635.

587 An overview of the status of the research into the origin of bipedalism is given by B. Wood, Four legs good, two legs better, *Nature* **363**, pp. 587–588, 17 June 1983. Cited on page 588.

588 More about the connection between entropy and computers can be found in the classic paper by R. Landauer, Irreversibility and heat generation in the computing process, *IBM Journal of Research and Development* **5**, pp. 183–191, 1961, and in C.H. Bennett & R. Landauer, The fundamental physical limits of computation, *Scientific American* **253**, pp. 48–56, 1985. Cited on page 591.

589 W.H. Zurek, Thermodynamic cost of computation, algorithmic complexity and the information metric, *Nature* **341**, pp. 119–124, 14 August 1989. Cited on page 591.

590 L. Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, *Zeitschrift für Physik* **53**, p. 840, 1929. This classic paper can also be found in English translation in the collected works by Leo Szilard. Cited on page 592.

591 J.J. Hopfield/ *Nature* **376**, pp. 33–36, 1995. This paper by one of the fathers of the field presents one possibility by which the *timing* of nerve signals, instead of the usually assumed firing frequency, could also carry information. Cited on page 592.

592 The details of the properties of the firing patterns of neurons are nicely described in the article by M. Mahowald & R. Douglas, A silicon neuron, *Nature* **354**, pp. 515–518, 19/26 December 1991, in which they show how to simulate a neuron's electrical behaviour using a silicon circuit. Cited on page 592.

593 The discussion whether the brain is or is not superior to a computer is nicely summarised by G. Vollmer, Algorithmen, Gehirne, Computer – Was sie können und was sie nicht können, Teil I und Teil II, *Naturwissenschaften* **78**, p. 481, 1991, and **78**, pp. 533–542, 1991. Cited on page 593.

594 However, the language with the largest available dictionary is Dutch, with the 40 volumes of the Wordenboek der Nederlandsche Taal, which appeared between 1864 and 1998. It has almost 400 000 entries. Cited on page 595.

595 The list and the remark on discovery on concepts is due to a personal communication from Anna Wierzbicka. A longer list is published in her book *Semantics, Primes and Universals*, Oxford University Press, 1996. Cited on pages 596 and 613.

596 W.S. Hatcher, *Foundations of Mathematics*, W.B. Saunders Co., Philadelphia, 1968. There is also the article by P. J. Cohen & R. Hersch, Non-Cantorian set theory, *Scientific American* **217**, pp. 104–116, 1967. Cohen was the mathematician who in 1963 proved that the negation of the continuum hypothesis could be added to the axioms of set theory and still produce a consistent theory; he calls such sets *non-Cantorian*. Cited on page 600.

597 See the beautiful article by I. Stewart, Fair shares for all, *New Scientist*, pp. 42–46, 17 June 1995. Cited on page 600.

598 The proof of the independence of the continuum hypothesis came in two parts. First, Kurt Gödel proved in 1940 that an axiom can be consistently added to zfc set theory so that the continuum hypothesis is correct. Then, in 1963, Paul Cohen proved that an axiom can be consistently added to zfc set theory so that the continuum hypothesis is false. Cited on page 602.

599 Rudy Rucker, *Infinity and the Mind – the Science and Philosophy of the Infinite*, Bantam, Toronto, 1983. Cited on page 602.

600 This general division of mathematics is nicely explained in the text by Pierre Basieux, *Die Architektur der Mathematik – Denken in Strukturen*, Rororo, 2000. Cited on page 604.

601 The issue is treated in Thomas Aquinas, *Summa Theologica*, in question 52 of the first part. The complete text, several thousand pages, can be found on the http://www.newadvent.org website. We come back to it in the part on quantum theory, in the section on the Pauli exclusion principle. It seems that the whole question goes back to Peter (the) Lombard, *Liber Sententiarum c.* 1150. Cited on page 605.

Page 732

602 I. Stewart, Daisy, daisy, give me your answer, do, *Scientific American*, pp. 76–79, January 1995. This pedagogical article explains how the growth of plants usually leads to flowers whose number of petals is from the Fibonacci series 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, etc. Deviations from this ideal case are also explained. The original work are two articles by S. Douady & Y. Couder, La physique des spirales végétales, *La Recherche* **24**, pp. 26–36, 1993, and Phyllotaxis as a self-organised growth process, in *Growth Patterns in Physical Sci-*

ences and Biology, edited by J.M. Garcia-Ruiz & al., Plenum Press, 1993. Cited on page 605.

603 H. Davson, *The Eye*, Academic Press, 1962. Cited on page 605.

604 See the http://akbar.marlboro.edu/~mahoney/cube/NxN.txt website. Cited on page 606.

605 An introduction to the surreal numbers is given by the article by Polly Shulman, Infinity plus one, and other surreal numbers, *Discover*, pp. 99–105, December 1995. There is also the text by D. Knuth, *Surreal Numbers: How two ex-Students Turned on to Pure Mathematics and Found Total Happiness*, Addison Wesley, 1974, or http://www-cs-faculty.stanford.edu/~knuth/sn.html. The usually quoted references on the topic include John H. Conway, *On Numbers and Games*, Academic Press, 1976, E.R. Berlekamp, J.H. Conway & R.K. Guy, *Winning Ways for Your Mathematical Plays, Volume I: Games in General*, Academic Press, 1982, and H. Gonshor, *An Introduction to Surreal Numbers*, Cambridge University Press, 1986. Cited on pages 606 and 608.

606 This beautiful problem is discussed by Ian Stewart, A bundling fool beats the wrap, *Scientific American,* pp. 109–111, June 1993. In four dimensions, the answer is known to lie somewhere between 50 000 and 100 000, whereas the five-dimensional answer is conjectured to be 'never'. Cited on page 608.

607 A. Pais, *Niels Bohr's Times: in Physics, Philosophy and Polity*, Oxford University Press, 1991, page 176. Cited on page 609.

608 Eugene Wigner, *Symmetries and Reflections*, Indiana University Press, 1962. Cited on page 609.

609 Göran Wikell, The layout of digits on pushbutton telephones – a review of the literature, *Tele* 34, pp. 34–42, 1982. Cited on page 611.

610 About the different aspects of falsifiability of general statements it is commonplace to cite the work by the epitemologist Karl Popper (1902–1994), especially his long and boring book *Logik der Forschung*, first published in 1934. The reason for this boredom is that Popper's work is simply a warming-up of Pierre Duhem's ideas. Cited on page 616.

611 For a good way of making blood that liquefies, see L. Garlaschelli, F. Ramaccini & S. Della Scala, Working bloody miracles, *Nature* 353, p. 507, 1991. The *Grand dictionnaire universel du XIXe siècle*, by Pierre Larousse, also contains a recipe; it was again shown to the public in the 1980s by Henri Broch. A wonderful and classic text is Harry Houdini, *Miracle Mongers and their Methods*, Prometheus Books, Buffalo, 1981. The original, written in 1920, by the world famous magician named 'The Great Houdini', is also available on the http://etext.lib.virginia.edu/toc/modeng/public/HouMirM.html website. The milk drinking Indian statues were common around the world in 1994 and 1995. About healers, see James Randi, *Flim-flam!*, Prometheus Books, Buffalo, New York, 1987, and the exquisite book by Hans Conrad Zander, *Warum ich Jesus nicht leiden kann*, Rowohlt, 1994. Cited on page 617.

612 Roger Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe*, Jonathan Cape, 2004, page 378. Cited on page 632.

613 John Horgan, *The End of Science – Facing the Limits of Knowledge in the Twilight of the Scientific Age*, Broadway Books, 1997, pp. 31–32, and chapter 2, note 2. Cited on page 619.

614 For an opinion completely contrary to the one described here, see the book by Gregory J. Chaitin, *The Limits of Mathematics*, Springer Verlag, 1997, which can also be found on the author's website http://www.cs.auckland.ac.nz/CDMTCS/chaitin/lm.html, along with his other works. Chaitin has devoted most of his life to the questions discussed in the section, especially on computability. Cited on page 620.

615 See the book by J. Barwise & J. Etchemendy, *The Liar*, Oxford University Press, New York, 1987. Cited on page 620.

616 This definition (statement 4.11) and many other statements about science are in the beautiful and rightly famous text by Ludwig Wittgenstein, *Tractatus logico-philosophicus*, Edition Suhrkamp, 1963. It gives a condensed summary of the basis of science, thought and language in a collection of highly structured and numbered sentences. Cited on pages 621 and 626.

617 See M. Dresden, The Klopsteg memorial lecture, *American Journal of Physics* **66**, pp. 468–482, 1998. Cited on page 622.

618 Well-known books are e.g. Friedrich Kohlrausch, *Praktische Physik*, Teubner, 24. Auflage, 1996. Cited on page 622.

619 Results are described e.g. in L. Bergmann & C. Schäfer, *Lehrbuch der Experimentalphysik*, Band I, II, III und IV, W. de Gruyter. Cited on page 622.

620 The origin of this incorrect attribution is the book by Gerhard Szczesny, *Brecht, Leben des Galilei – Dichtung und Wirklichkeit*, Ullstein, Berlin 1966, p. 6. The statement has never been made by Galilei; this issue has been discussed at length in specialist circles, e.g. by F. Kleinert, "Messen was meßbar ist" - Über ein angebliches Galilei-Zitat, *Berichte zur Wissenschaftgeschichte* **11**, p. 221, 1988, or on the internet by Peter Jaencke. Cited on page 623.

621 Landolt-Börnstein, edited by K.-H. Hellwege & O. Madelung, *Zahlenwerte und Funktionen aus Naturwissenschaften und Technik*, Neue Serie, Springer Verlag, Berlin, 1984. This series of more than one hundred volumes contains all important observations in the domain of physics. Cited on page 622.

622 The strange and sometimes dangerous consequences of beliefs can be found e.g. in Martin Gardner, *Fads and Fallacies*, Dover, 1957, and in James Randi, *Faith Healers*, Prometheus Books, 1989. The million dollar prize for showing any paranormal or supernormal effect is available from his http://www.randi.org website. Cited on page 627.

623 See the nice collection of cranks on the http://www.crank.net web site. Cited on page 627.

624 As an example, the opposite view on the emergence of properties is strongly defended by P. Jensen, Particle physics and our everyday world, *Physics Today* pp. 58–59, July 1998. Cited on page 628.

625 See page 133 of the bibliography by John Bowlby, *Charles Darwin*, Pimlico, 1991. Cited on page 629.

626 John Horgan, *The End of Science – Facing the Limits of Knowledge in the Twilight of Scientific Age*, Addison Wesley, 1996. Cited on page 632.

627 A beautiful introduction to Greek philosophy is Eduard Zeller, *Outlines of the History of Greek Philosophy*, Dover, 1980, a reprint of a book published in 1928. Among others, it gives a clear exposition of the philosophy of Democritus and the other presocratics. Cited on page 633.

628 The famous quote is found at the beginning of chapter XI, 'The Physical Universe', in Arthur Eddington, *The Philosophy of Physical Science*, Cambridge, 1939. Cited on page 633.

629 Giuseppe Fumagalli, *Chi l'ha detto?*, Hoepli, Milano, 1983. Cited on page 635.

630 See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 653, 1991. Cited on page 635.

631 For a beautiful text on fractals, see the footnote on page 52. Cited on page 637.

632 As has been pointed out by René Descartes. Cited on page 640.

633 The famous carbon ^{12}C resonance was found by Willy Fowler, as described in E. Margaret Burbridge, G.R. Burbridge, W.A. Fowler & F. Hoyle, Synthesis of the elements in stars, *Reviews of Modern Physics* **29**, pp. 547–560, 1957. Cited on page 642.

634 An extensive overview of the topic is given in the thick book by John D. Barrow & Frank J. Tipler, *The Anthropic Cosmological Principle*, Oxford University Press, 1986. The term itself is due to Brandon Carter, who coined it in 1973 and presented it in a symposium devoted to the 500th anniversary of Nicolaus Copernicus. For more literature, see Yuri I. Balashov, Resource Letter AP-1: the anthropic principle, *American Journal of Physics* **59**, pp. 1069–1076, 1991. Cited on page 642.

635 Voltaire, *Candide ou l'optimisme*, 1759. See the footnote on page 166. The book is so good that it was still being seized by the US customs in 1930, and the US post office refused to transport it as late as 1944. For more details, search for 'banned books online' on the worldwide web. Cited on page 643.

636 The number of books on consciousness is large and the contents not always interesting, and often not based on fact, as shown by Karl R. Popper & John Eccles, *The Self and its Brain – an Argument for Interactionism*, Rutledge, 1993. Cited on page 644.

637 See e.g. the *Encyclopedia Britannica*, Macropaedia, in the entry on animal behaviour. Cited on page 644.

638 A straight and informative introduction to the work and ideas of Joseph Beuys (in German) is by Renate Georgi, Joseph Beuys, RAAbits Kunst, Raabe Fachverlag, September 1996. Cited on page 645.

639 Two studies, one by R.P. Ebstein & al., Dopamine D4 receptor (D4DR) exon III polymorphism associated with human personality trait of novelty seeking, *Nature Genetics* **12**, pp. 78–80, January 1996, and another study by J. Benjamin & al., Population and familial association between the D4 dopamine receptor gene and measures of novelty seeking, *Nature Genetics* **12**, pp. 81–84, January 1996, found that people with a special form of the D4 dopamine receptor gene, or D4DR, are more prone to novelty seeking than people with the usual form. The D4DR gene regulates the formation of dopamine receptors, a chemical messenger in the brain that has been a candidate for some time for a substance involved in the propensity for novelty seeking. Cited on page 645.

640 More can be found in Jacques Hadamard, *The Mathematician's Mind – The Psychology of Invention in the Mathematical Field*, Princeton Science Library, 1996. Cited on page 645.

641 For an exploration of 'inner' motions, see the beautiful text by Richard Schwartz, *Internal Family Systems Therapy*, The Guilford Press, 1995. No citations.

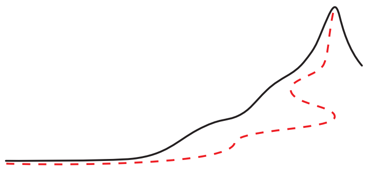
642 A clear overview of philosophy of science, often called *epistemology*, without unnecessary detail, is given by Robert Blanché, L'Epistémologie, Presses Universitaires de France, 1972. No citations.

643 An introduction to neurology is Joseph Ledoux, *Synaptic Self: How Our Brains Become Who We Are*, Viking Press, 2002. No citations.

644 An introduction to computer science is given in J. Glenn Brookshear, *Computer Science, An Overview*, 6th edition, Addison Wesley, 2000, or in Rick Decker & Stuart Hirshfield, *The Analytical Engine: An Introduction to Computer Science Using the Internet*, Brooks/Cole Publishers, 1998. No citations.

645 Alfred Tarski, *Introduction to Modern Logic*, Dover, 1946. See also the famous children's book by the mathematician and photographer Lewis Carroll, *Alice in Wonderland*. No citations.

646 For example, one needs the courage to face envy. About this topic see the classic text by Helmut Schoeck, *Der Neid*, 1966, published in English as *Envy: A Theory of Social Behavior*, 1969. It is the standard work in the field. Cited on page 647.

Second Part 

Quantum Theory:
What Is Matter?
What Are Interactions?

Where the existence of a minimal change is deduced,
implying that motion is fuzzy,
that matter is not permanent,
that boxes are never tight,
that matter is composed of elementary units
and that light and interactions are streams of particles,
thus explaining why antimatter exists,
why the floor does not fall but keeps on carrying us,
why particles are unlike gloves,
why empty space pulls mirrors together
and why the stars shine.

Chapter V

Quanta of Light and Matter

19. Minimum action – quantum theory for poets and lawyers

Natura [in operationibus suis] non facit saltus.*
Ref. 647 15th century

Escalating Motion Mountain up to this point, we completed three legs. We first encountered Galileo's mechanics, the description of motion for kids, then Einstein's relativity, the description of motion for science fiction enthusiasts, and finally Maxwell's electrodynamics, the description of motion valuable to craftsmen and businessmen.

These three classical descriptions of motion are impressive, beautiful and useful. However, they also have a small problem: they are wrong. The reason is simple: none of them describes *life*. Whenever we observe a flower, we enjoy its bright colours, its wild smell, its soft and delicate shape or the fine details of its symmetry. None of the three classical descriptions can explain any of these properties; neither do they explain the impression they make on our senses. Classical physics can describe them partly, but it cannot explain their origins. For an explanation, we need *quantum theory*. In fact we will discover that *every type of pleasure* in life is an example of quantum motion. Just try; take any example of a pleasant situation, such as a beautiful evening sky, a waterfall, a caress or a happy child. Classical physics is not
Challenge 1122 n able to explain it: the involved colours, shapes and sizes remain mysterious.

Figure of orchidea to be inserted

Figure 272 An example of a quantum system

In the beginning of physics this limitation was not seen as a shortcoming: in those times neither senses nor material properties were imagined to be related to motion. And of course, in older times the study of pleasure was not deemed a serious topic of invest-
Page 231 igation for a respectable researcher. However, in the meantime we know that the senses of touch, smell and sight are first of all *detectors of motion*. Without motion, no senses! In addition, all detectors are built of matter. In the chapter on electromagnetism we started to understand that all properties of matter are due to motion of charged constituents. Density, stiffness, colour and all other material properties result from the electromagnetic behaviour of the Lego bricks of matter, namely the molecules, the atoms and the

* Nature [in its workings] makes no jumps.

electrons. Thus, also matter properties are consequences of motion. In addition, we saw that these tiny constituents are *not* correctly described by classical electrodynamics. We
Page 565 even found that light itself behaves unclassically. Therefore the inability of classical physics to describe matter and the senses is indeed due to its intrinsic limitations.

In fact, every failure of classical physics can be traced back to a single, fundamental
Ref. 648 discovery made in 1899 by Max Planck:*

▷ *In nature, actions smaller than the value* $\hbar/2 = 0.53 \cdot 10^{-34}$ Js *are not observed.*

All experiments trying to do so invariably fail. In other words, in nature there is always some action – like in a good movie. This existence of a minimal action, the *quantum prin-*
Challenge 1123 n *ciple*, is in *full* contrast with classical physics. (Why?) However, it has passed the largest imaginable number of experimental confirmations, many of which we will encounter in this second part of our mountain ascent. Planck discovered the principle when study-
Page 565 ing the properties of incandescent light, i.e. the light emanating from hot bodies. But the quantum principle also applies to motion of matter, and even, as we will see later, to motion of space-time. By the way, the factor 1/2 results from the historical accidents in the definition of the constant $\hbar$, which is read as 'eitch-bar'. Despite the missing factor, the constant $\hbar$ is called the *quantum of action* or also, after its discoverer, *(reduced) Planck's constant*.

The quantum principle states that no experiment whatsoever can measure an action value smaller than $\hbar/2$. For a long time, even Einstein tried to devise experiments to overcome the limit. But he failed: nature does not allow it.

Interestingly, since action in physics, like action in the movie industry, is a way to measure the *change* occurring in a system, a minimum action implies that *there is a minimum change in nature.* The quantum of action thus would be better named the *quantum of change*. Comparing two observations, there always is change. (What is called 'change' in everyday life is often called 'change of state' by physicists; the content is the same.) Before we cite all the experiments confirming this statement, we give an introduction to some of its more surprising consequences.

Since action measures change, a minimum observable action means that two subsequent observations of the same system always differ by at least $\hbar/2$. In every system, there is always *something* happening. As a consequence, in nature *there is no rest.*
Page 656 Everything moves, all the time, at least a little bit. Natura facit saltus. True, it is only a tiny bit, as the value of $\hbar/2$ is so small. For example, the quantum of action implies that in a mountain, a system at rest if there is any, all atoms and all electrons are continuously buzzing around. Rest can be observed only macroscopically, and only as a long time or many particle average.

Since there is a minimum action for all observers, and since there is no rest, in nature *there is no perfectly straight and no perfectly uniform motion*. Forget all you have learned

* This somewhat unconventional, but useful didactic approach is due to Niels Bohr. (Bohr though did not know about the factor 1/2 when he propagated it. Nowadays, the approach is almost never found in the
Ref. 649 literature; it might be used in a teaching text for the first time here.

About Max Planck and his accomplishments, see the footnote on page 565. In fact, the cited quantum principle is a simplification; the constant originally introduced by Planck was the (unreduced) constant $h = 2\pi\hbar$. The factors 2π and 1/2 leading to the final quantum principle were found somewhat later, by other researchers.

so far. Every object moves in straight and uniform motion only approximately, and only when observed over long distances or long times. We will see later that the more massive
Challenge 1124 ny the object is, the better the approximation is. Can you confirm this? As a consequence, macroscopic observers can still talk about space-time symmetries. *Special* relativity can thus easily be reconciled with quantum theory.

Obviously, also free fall, i.e. motion along geodesics, exists only as a long time average. In this sense, *general* relativity, being based on the existence of freely falling observers, cannot be correct when actions of the order of $\hbar$ are involved. Indeed, the reconciliation of the quantum principle with *general* relativity – and thus with curved space – is a big challenge. The issues are so mind-shattering that the topic forms a separate, third part of this mountain ascent.

Have you ever wondered why leaves are green? Probably you know that they are green because they absorb blue light, of short wavelengths, and red light, of long wavelengths, and let green, medium wavelength light be reflected. How can a system filter out the small and the large, and let the middle pass through? To do so, leaves must somehow *measure* the wavelength. But we have seen that classical physics does not allow to measure length or time intervals, as any measurement requires a measurement unit, and classical physics does not allow to define units for them. On the other hand, it takes only a few lines to confirm that with the help of the quantum of action $\hbar$ (and the Boltzmann constant k, which Planck discovered at the same time), fundamental measurement units of *all* measurable quantities can be defined, including length and thus wavelength. Can you find a
Challenge 1125 n combination of c, G and $\hbar$ giving a length? It only will take a few minutes. When Planck found the combination, he was happy like a child; he knew straight away that he had made a fundamental discovery, even though in 1899 quantum theory did not exist yet. He even told his seven year old son Erwin abut it, while walking with him through the forests around Berlin. Planck explained to his son that he had probably made a discovery as important as universal gravity. Indeed, Planck knew that he had found the key to understanding most of the effects which were unexplained so far. In particular, without the quantum of action, colours would not exist. Every colour is a quantum effect.*

Planck realized that the quantum of action allows to understand the *size* of all things. With the quantum of action, it was finally possible to answer the question on the max-
Page 203 imum size of mountains, of trees and of humans. Planck knew that the quantum of action confirmed the answer Galileo had deduced already long before him: sizes are due to fundamental, minimal scales in nature. The way the quantum of action allows to understand the sizes of physical systems will be uncovered step by step in the following.

The size of objects is related to the size of atoms; in turn, the *size of atoms* is a *direct consequence* of the quantum of action. Can you deduce an approximation for the size of atoms, knowing that it is given by the motion of electrons of mass m_e and charge e,
Challenge 1127 n constrained by the quantum of action? This connection, a simple formula, was discovered in 1910 by Arthur Erich Haas, 15 years before quantum theory was formulated; at the time, everybody made fun of him. Nowadays, the expression is found in all textbooks.**

Challenge 1126 n * It is also possible to define all units using c, G and e, the electron charge. Why is this not satisfactory?

** Before the discovery of $\hbar$, the only simple length scale for the electron was the combination $e^2/(4\pi\varepsilon_0 mc^2) \approx 3$ fm; this value is ten thousand times smaller than an atom.

By determining the size of atoms, the quantum of action has an important consequence: Gulliver's travels are impossible. There are no tiny people and no giant ones. Classically, nothing speaks against the idea; but the quantum of action does. Can you provide the detailed argument? Challenge 1128 n

But if rest does not exist, how can *shapes* exist? Any shape, also that of a flower, is the result of body parts remaining *at rest* with respect to each other. Now, all shapes result from the interactions of matter constituents, as shown most clearly in the shape of molecules. But how can a molecule, such as the water molecule H_2O, have a shape? In fact, it does not have a *fixed* shape, but its shape fluctuates, as expected from the quantum of action. Despite the fluctuations it does have an *average* shape, because different angles and distances correspond to different energies. And again, these average length and angle values only result because the quantum of action leads to fundamental length scales in nature. Without the quantum of action, there would be *no* shapes in nature.

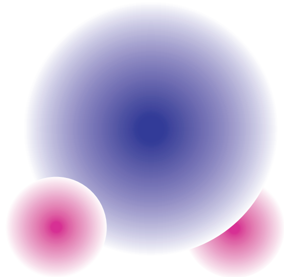


Figure 273 An artistic impression of a water molecule

As we will discover shortly, quantum effects surround us from all sides. However, since the minimum action is so small, its effects *on motion* appear mostly, but not exclusively, in *microscopic* systems. The study of such systems has been called *quantum mechanics* by Max Born, one of the main figures of the field.* Later on, the term *quantum theory* became more popular. In any case, quantum physics is the description of microscopic motion. But when is quantum theory necessary? Table 54 shows that all processes on atomic and molecular scale, including biological and chemical ones, involve action values near the quantum of action. So do processes of light emission and absorption. All these phenomena can be described *only* with quantum theory.

The term 'quantum' theory, by the way, does not mean that all measurement values are *multiples* of a smallest one; this is correct only in certain cases. Quantum theory means the existence of *minimum* measurable values, precisely in the way that Galileo already speculated about in the seventeenth century. As mentioned in detail earlier on, it was Galileo's insistence on these 'piccolissimi quanti' that got him condemned to lifelong imprisonment, and not, as is usually told, his ideas on the motion of the Earth. Of course, we will discover that only the idea of a smallest change leads to a precise and accurate description of nature. Page 203

Table 54 also shows that the term 'microscopic' has a different meaning for a physicist and for a biologist. For a biologist, a system is microscopic if it requires a *microscope* for

* Max Born (b. 1882 Breslau, d. 1970 Göttingen) first studied mathematics, then turned to physics. Professor in Göttingen, he made the city one of the world centres of physics. He developed quantum mechanics with his assistants Werner Heisenberg and Pascual Jordan, then applied it to scattering, to solid state physics, to optics and to liquids. He was the physicist who first understood that the state function describes a probability amplitude. He is one of the authors of the famous Born & Wolf textbook on optics; it still remains the main book of the field. He attracted to Göttingen the most brilliant talents of the time, receiving as visitors Hund, Pauli, Nordheim, Oppenheimer, Goeppert–Mayer, Condon, Pauling, Fock, Frenkel, Tamm, Dirac, Mott, Klein, Heitler, London, von Neumann, Teller, Wigner and dozens of others. Being Jewish, Max Born lost his job in 1933; he emigrated and became professor in Edinburgh, where he stayed for twenty years. Physics at Göttingen University never recovered from this loss. For his elucidation of the meaning of the wave function he received the 1954 Nobel prize in physics. Ref. 650

Table 54 Some small systems in motion and the observed action values for their changes

System & change	Action	Motion
Light		
Smallest amount of light absorbed by a coloured surface	$1\,\hbar$	quantum
Smallest hit when light reflects from mirror	$2\,\hbar$	quantum
Smallest visible amount of light	*c.* $5\,\hbar$	quantum
Smallest amount of light absorbed in flower petal	*c.* $1\,\hbar$	quantum
Blackening of photographic film	*c.* $3\,\hbar$	quantum
Photographic flash	*c.* $10^{17}\,\hbar$	classical
Electricity		
Electron ejected from atom	*c.* $1-2\,\hbar$	quantum
Electron added to molecule	*c.* $1-2\,\hbar$	quantum
Electron extracted from metal	*c.* $1-2\,\hbar$	quantum
Electron motion inside microprocessor	*c.* $2-6\,\hbar$	quantum
Signal transport in nerves, from one molecule to the next	*c.* $5\,\hbar$	quantum
Current flow in lighting bolt	*c.* $10^{38}\,\hbar$	classical
Materials science		
Tearing apart two neighbouring iron atoms	*c.* $1-2\,\hbar$	quantum
Breaking a steel bar	*c.* $10^{35}\,\hbar$	classical
Basic process in superconductivity	$1\,\hbar$	quantum
Basic process in transistors	$1\,\hbar$	quantum
Basic process in magnetic effects	$1\,\hbar$	quantum
Chemistry		
Atom collisions in liquids at room temperature	*c.* $1\,\hbar$	quantum
Shape oscillation of water molecule	*c.* $1-5\,\hbar$	quantum
Shape change of molecule, e.g. in chemical reaction	*c.* $1-5\,\hbar$	quantum
Single chemical reaction curling a hair	*c.* $2-6\,\hbar$	quantum
Tearing apart two mozzarella molecules	*c.* $300\,\hbar$	quantum
Smelling one molecule	*c.* $10\,\hbar$	quantum
Burning fuel in a cylinder in an average car engine explosion	*c.* $10^{37}\,\hbar$	classical
Life		
Air molecule hitting ear drum	*c.* $2\,\hbar$	quantum
Smallest sound signal detectable by the ear	challenge	classical
DNA duplication step in cell division	*c.* $100\,\hbar$	quantum
Ovule fecundation	*c.* $10^{14}\,\hbar$	classical
Smallest step in molecular motor	*c.* $5\,\hbar$	quantum
Sperm motion by one cell length	*c.* $10^{15}\,\hbar$	classical
Cell division	*c.* $10^{19}\,\hbar$	classical
Fruit fly's wing beat	*c.* $10^{24}\,\hbar$	classical
Person walking one body length	*c.* $2\cdot10^{36}\,\hbar$	classical
Nuclei and stars		
Nuclear fusion reaction in star	*c.* $1-5\,\hbar$	quantum
Particle collision in accelerator	*c.* $1\,\hbar$	quantum
Explosion of gamma ray burster	*c.* $10^{80}\,\hbar$	classical

its observation. For a physicist however, a system is microscopic if its characteristic action is of the order of the quantum of action. In short, for a physicist, a system is microscopic if it is *not* visible in a (light) microscope. To increase the confusion, some quantum physicists nowadays call their own class of microscopic systems 'mesoscopic,' whereas many classical, macroscopic systems are now called 'nanoscopic'. Both names mainly help to attract attention and funding.

There is another way to characterize the difference between a microscopic or quantum system on one side and a macroscopic or classical system on the other. A minimum action implies that the difference of action values S between two successive observations of the same system, spaced by a time Δt, is limited. Therefore one has

$$S(t+\Delta t)-S(t)=(E+\Delta E)(t+\Delta t)-Et=E\Delta t+t\Delta E+\Delta E\Delta t\geqslant\frac{\hbar}{2}\,. \qquad (472)$$

Now the value of the energy E and of the time t – but not that of ΔE or of Δt – can be set to zero if we choose a suitable observer; thus the existence of a quantum of action implies that in any system the evolution is constrained by

$$\Delta E\Delta t\geqslant\frac{\hbar}{2}\,, \qquad (473)$$

where E is the energy of the system and t its age. In other words, ΔE is the change of energy and Δt the time between two successive observations. By a similar reasoning we find that for any system the position and momentum values are constrained by

Challenge 1129 ny

$$\Delta x\Delta p\geqslant\frac{\hbar}{2}\,, \qquad (474)$$

where Δp is the indeterminacy in momentum and Δp the indeterminacy in position. These two famous relations were called *indeterminacy relations* by their discoverer, Werner Heisenberg.* The name is often incorrectly translated into English as 'uncertainty relations'. However, this latter name is wrong: the quantities are not uncertain, but *undetermined*. Due to the quantum of action, system observables have *no* definite value.

* One often hears the myth that the indeterminacy relation for energy and time has another weight than the one for momentum and position. That is wrong; it is a myth propagated by the older generation of physicists. This myth survived through many textbooks for over 70 years; just forget it, as it is incorrect. It is essential to remember that all four quantities appearing in the inequalities are quantities describing the *internal* properties of the system. In particular, it means that t is some time variable deduced from changes observed *inside* the system and *not* the external time coordinate measured by an outside clock, in the same way that the position x is *not* the external space coordinate, but the position characterizing the system.

Ref. 651

Werner Heisenberg (1901–1976) was an important German theoretical physicist and an excellent table tennis and tennis player. In 1925, as a young man, he developed, with some help by Max Born and Pascual Jordan, the first version of quantum theory; from it he deduced the indeterminacy relations. For these achievements he received the Nobel prize for physics in 1932. He also worked on nuclear physics and on turbulence. During the second world war, he worked in the German nuclear fission program. After the war, he published several successful books on philosophical questions in physics and he unsuccessfully tried, with some half-hearted help by Wolfgang Pauli, to find a unified description of nature based on quantum theory, the 'world formula'.

There is *no* way to ascribe a precise value to momentum, position and other observables of a quantum system.

Any system whose indeterminacy is of the order of $\hbar$ is a quantum system; if the indeterminacy product is much larger, the system is classical, and classical physics is sufficient for its description. In other words, even though classical physics assumes that there are *no* measurement indeterminacies in nature, a system is classical only if its indeterminacies are *large* compared to the minimum possible ones. As a result, quantum theory is necessary in all those cases in which one tries to measure some quantity as precisely as possible.

The indeterminacy relations again show that *motion cannot be observed to infinite precision.* In other words, the microscopic world is *fuzzy*. This strange result has many important and many curious consequences. For example, if motion cannot be observed with infinite precision, the very concept of motion needs to be used with great care, as it cannot be applied in certain situations. In a sense, the rest of our quest is an exploration of the implications of this result. In fact, as long as space-time is *flat*, it turns out that we *can* keep motion as a concept describing observations, provided we remain aware of the limitations of the quantum principle.

In particular, the quantum of action implies short-time deviations from energy, momentum and angular momentum conservation in microscopic systems. Now, in the first Page 141 part of our mountain ascent we realized that any type of nonconservation implies the existence of surprises in nature. Well, here are some of them.

Since uniform motion does not exist in the precise meaning of the term, a system moving in one dimension only, such as the hand of a clock, always has a possibility to move a bit in the opposite direction, thus leading to incorrect readings. Indeed, quantum theory predicts that *clocks have limits*, and that perfect clocks do not exist. In fact, quantum theory implies that strictly speaking, one-dimensional motion does not exist.

Challenge 1130 ny Obviously, the limitations apply also to meter bars. Thus the quantum of action is responsible on one hand that the possibility to perform measurements exists, and on the other hand for the limitations of measurements.

In addition, it follows from the quantum of action that any observer must be *large* to be inertial or freely falling, as only large systems approximate inertial motion. *An observer cannot be microscopic*. If humans were not macroscopic, they could neither observe nor study motion.

Due to the finite accuracy with which microscopic motion can be observed, faster than light motion should be possible in the microscopic domain. Quantum theory thus predicts *tachyons*, at least over short time intervals. For the same reason, also *motion backwards in time* should be possible over microscopic times and distances. In short, a quantum of action implies the existence of microscopic time travel.

But there is more: the quantum of action implies that *there is no permanence in nature.* Imagine a moving car suddenly disappearing for ever. In such a situation neither momentum nor energy would be conserved. The action change for such a disappearance is large compared to $\hbar$, so that its observation would contradict even classical physics, as you Challenge 1131 ny might want to check. However, the quantum of action allows that a *microscopic* particle, such as an electron, disappears for a *short* time, provided it reappears afterwards.

The quantum of action also implies that *the vacuum is not empty*. If one looks at empty space twice in a row, the two observations being spaced by a tiny time interval, some energy will be observed the second time. If the time interval is short enough, due to the quantum of action, matter particles will be observed. Indeed, particles can appear anywhere from nowhere, and disappear just afterwards, as the action limit requires it. In other words, classical physics' idea of an *empty* vacuum is correct only when observed over *long* time scales. In summary, nature shows short time appearance and disappearance of matter.

The quantum of action implies that compass needles cannot work. If we look twice in a row at a compass needle or even at a house, we usually observe that they stay oriented in the same direction. But since physical action has the same unit as angular momentum, a minimum value for action also means a minimum value for angular momentum. Therefore, every macroscopic object has a minimum value for its rotation. In other words, quantum theory predicts that *in everyday life, everything rotates*. Lack of rotation exists only approximately, when observations are spaced by long time intervals.

Challenge 1132 ny

For microscopic systems, the situation is more involved. If their rotation angle *can* be observed, such as for molecules, they behave like macroscopic objects: their position and their orientation are fuzzy. But for those systems whose rotation angle *cannot* be observed, the quantum of action turns out to have somewhat different consequences. Their angular momentum is limited to values which are multiples of $\hbar/2$. As a result, all microscopic bound systems, such as molecules, atoms, or nuclei, contain rotational motion and rotating components.

But there is more to come. A minimum action implies that cages in zoos are dangerous and banks are not safe. A cage is a feature requiring a lot of energy to be overcome. Mathematically, the wall of a cage is an energy hill, similar to the one shown in Figure 274. If a particle on one side of the hill has momentum p, it is simple to show that the particle can be observed on the other side of the hill, at position Δx, even if its kinetic energy $p^2/2m$ is *smaller* than the height E of the hill. In everyday life this is impossible. But imagine that the missing momentum $\Delta p = \sqrt{2mE - p^2}$ to overcome the hill satisfies $\Delta x \Delta p \geqslant \hbar/2$. The quantum of action thus implies that a hill of width

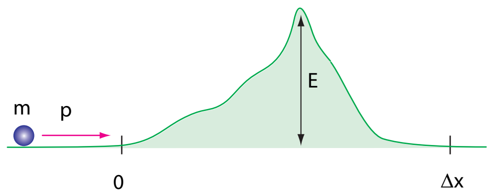


Figure 274 Hills are never high enough

$$\Delta x \leqslant \frac{\hbar/2}{\sqrt{2mE - p^2}} \tag{475}$$

is *not* an obstacle to a particle of mass m. But this is not all. Since the value of the particle momentum p is itself undetermined, a particle can overcome the hill even if the hill is *wider* than value (475), though the broader it is the smaller the probability is. As a result, any particle can overcome *any* obstacle. This effect, for obvious reasons, is called the *tunnelling effect*. In short, the minimum action principle implies that there are no safe boxes in nature. Due to tunnelling, *matter is not impenetrable*, in contrast to everyday, classical observation. Can you explain why lion cages work *despite* the quantum of action?

Challenge 1133 ny

By the way, the quantum of action also implies that a particle with a kinetic energy larger than the energy height of a hill can get reflected by the hill. Classically this is im-

Challenge 1134 ny possible. Can you explain the observation?

The minimum action principle also implies that book shelves are dangerous. Shelves are obstacles to motion. A book in a shelf is in the same situation as the mass in Figure 275; the mass is surrounded by energy hills hindering its escape to the outer, lower energy world. Now, due to the tunnelling effect, escape is always possible. The same picture applies to a branch of a tree, a nail in a wall, or to anything attached to anything else. Fixing things to each other is never for ever. We will find out that every example of light emission, even radioactivity, results from this effect. The quantum of action thus implies that *decay is part of nature*. In short, there are no stable excited systems in nature. For the same reason by the way, *no* memory can be perfect. (Can you confirm the deduction?) Note that decay often appears in everyday life, where it just has a different name: *breaking*. In fact, all cases in which something breaks require the quantum of action for their description. Obviously, the cause of breaking is often classical, but the *mechanism* of breaking is always quantum. Only objects that follow quantum theory can break.

Challenge 1135 ny

Ref. 652

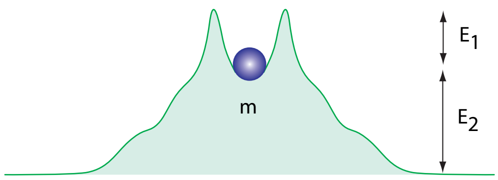


Figure 275 Leaving enclosures

Taking a more general view, also *aging* and *death* result from the quantum of action. Death, like aging, is a composition of breaking processes. Breaking is a form of decay, and is due to tunnelling. Death is thus a quantum process. Classically, death does not exist. Might this be the reason that so many believe in immortality or eternal youth?

Challenge 1136 ny

A minimum action also implies that matter cannot be continuous, but must be composed of smallest entities. Indeed, the flow of a truly continuous material would contradict the quantum principle. Can you give the precise argument? Of course, at this point of our adventure, the non-continuity of matter is no news any more. In addition, the quantum of action implies that even *radiation* cannot be continuous. As Albert Einstein stated clearly for the first time, light is made of quantum particles. More generally, the quantum of action implies that in nature *all flows and all waves are made of microscopic particles*. The term ‘microscopic’ or ‘quantum’ is essential, as such particles do *not* behave like little stones. We have already encountered several differences and we will encounter others shortly. For these reasons, microscopic particles should bear a special name; but all proposals, of which *quanton* is the most popular, have not caught on yet.

Challenge 1137 ny

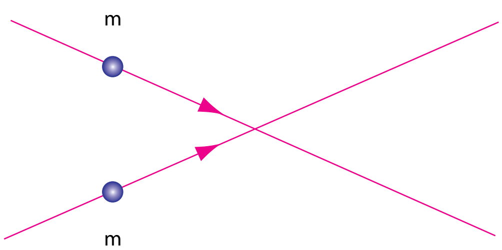


Figure 276 Identical objects with crossing paths

The quantum of action has several strange consequences for microscopic particles. Take two of them with the same mass and the same composition. Imagine that their paths cross and that at the crossing they approach each other to small distances, as shown in Figure 276. A minimum action implies that in such a situation, if the distance becomes small enough, the two particles can switch role without anybody being able to avoid or to ever notice it. For example, in a gas it is *impossible*, due to the quantum of action, to follow particles moving around and to say which particle is which. Can you confirm this

Challenge 1138 ny deduction and specify the conditions using the indeterminacy relations? In summary, in nature *it is impossible to distinguish identical particles.* Can you guess what happens in the case of light? Challenge 1139 n

But matter deserves still more attention. Imagine two particles, even two different ones, approaching each other to small distances, as shown in Figure 277. We know that if the approach distance gets small, things get fuzzy. Now, if something happens in that small domain in such a way that the resulting outgoing products have the same total momentum and energy as the incoming ones, the minimum action principle makes such processes possible. Indeed, ruling out such processes would imply that arbitrary small actions could be observed, thus eliminating nature's fuzziness, as you might want to check by yourself. In short, *a minimum action allows transformation of matter*. One also says that the quantum of action allows particle *reactions*. In fact, we will discover that *all* kinds of reactions in nature, including chemical and nuclear ones, are only due to the existence of the quantum of action.

Challenge 1140 ny

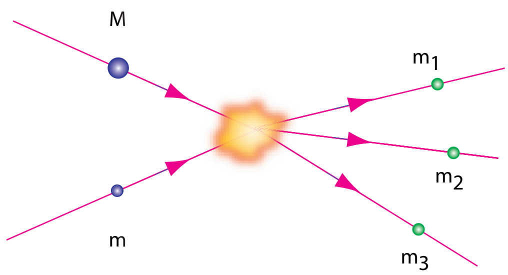


Figure 277 Transformation through reaction

But there is more. Due to the indeterminacy relations, it is impossible to give a definite value to both the momentum and the position of a particle. Obviously, this is also impossible for all the components of a measurement set-up or an observer. This implies that initial conditions – both for a system *and* for the measurement set-up – cannot be exactly duplicated. A minimum action thus implies that whenever an experiment on a microscopic system is performed twice, the outcome will be *different*. The result would be the same only if both the system and the observer would be in exactly the same condition in both situations. This turns out to be impossible, both due to the second principle of thermodynamics and due to the quantum principle. Therefore, *microscopic systems behave randomly*. Obviously, there will be some *average* outcome; nevertheless, microscopic observations are probabilistic. Albert Einstein found this conclusion of quantum theory the most difficult to swallow, as this randomness implies that the behaviour of quantum systems is strikingly different from that of classical systems. But the conclusion is unavoidable: nature behaves randomly.

A good example is given by trains. Einstein used trains to develop and explain relativity. But trains are also important for quantum physics. Everybody knows that one can use a train window to look either at the *outside* landscape or, by concentrating on the reflected image, to observe some interesting person *inside* the carriage. In other words, glass reflects some of the light particles and lets some others pass through. More precisely, glass reflects a random selection of light particles, yet with constant average. Partial reflection is thus similar to the tunnelling effect. Indeed, the partial reflection of glass for photons is a result of the quantum of action. Again, the average situation *can be described* by classical physics, but the precise amount of partial reflection *cannot be explained* without quantum theory. Without the quantum of action, train trips would be much more boring.

Finally, the quantum of action implies a famous result about the *path* of particles. If a particle travels from a point to another, there is no way to say which path it has taken

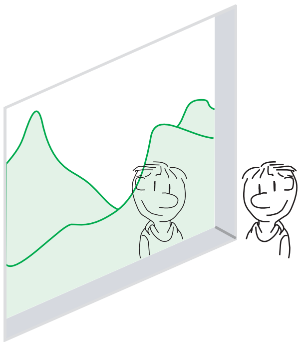


Figure 278 How do train windows manage to show two superimposed images?

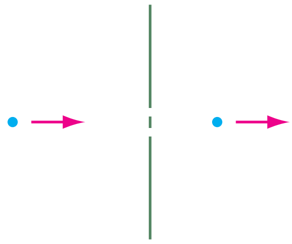


Figure 279 A particle and a screen with two nearby slits

in between. Indeed, in order to distinguish among the two possible, but only slightly different paths, actions smaller than $\hbar/2$ would have to be measured. In particular, if a particle is sent through a screen with two sufficiently close slits, it is impossible to say through which slit the particle passed to the other side. The impossibility is fundamental. As we will find out soon, this impossibility leads to particle interference and the wave behaviour of particles.

We will also discover that the quantum of action is the origin for the importance of the action observable in classical physics. In fact, the existence of a *minimal* action is the reason for the *least* action principle of classical physics.

In computer science, a smallest change is called a 'bit'. The existence of a smallest change in nature thus means that computer language or information science can be used to describe nature, and in particular quantum theory. However, computer language can describe only the software side; the hardware side of nature is also required. The hardware of nature enters the description whenever the actual value $\hbar$ of the quantum of action must be introduced.

Exploring the analogy between nature and information science, we will discover that the quantum of action implies that macroscopic physical systems cannot be copied or 'cloned', as quantum theorists like to say. Nature does not allow to copy objects. Copying machines do not exist. The quantum of action makes it impossible to gather and use all information in a way that allows to produce a perfect copy. As a result, we will deduce that the precise order in which measurements are performed does play a role in experiments. When the order is important, physicists speak of lack of 'commutation'. In short physical observables do not commute.

We will also find out that the quantum of action implies that systems are not always independent, but can be 'entangled'. This term, introduced by Erwin Schrödinger, describes the most absurd consequences of quantum theory. Entanglement makes everything in nature connected to everything else. Entanglement produces effects which look (but are not) faster than light. Entanglement produces a (fake) form of non-locality. Entanglement also implies that trustworthy communication does not exist.

Page 746

Ref. 653

Don't all these deductions look wrong or at least crazy? In fact, if you or your lawyer made any of these statements in court, maybe even under oath, you would be likely to end up in prison! However, all above statements are correct, as they are all confirmed by experiment. And the surprises are by far not finished. You might have noticed that, in the preceding examples, no situation related to electricity, to the nuclear interactions or to gravity was included. In these domains the surprises are even more astonishing; the observation of antimatter, of electric current flow without resistance, of the motion inside muscles, of vacuum energy, of nuclear reactions in stars and maybe soon of boiling empty space will fascinate you as much as they have fascinated and still fascinate thousands of researchers.

Challenge 1141 ny In particular, the consequences of the quantum of action on the early universe are simply mind-boggling. Just try to explore for yourself its consequences for the big bang. Together, all these topics will lead us towards the top of Motion Mountain. The topics are so strange, so incredible and at the same time so numerous that quantum physics can be rightly called the description of motion for *crazy* scientists. In a sense, this is the generalization of the previous definition, when we called quantum physics the description of motion related to pleasure.

Sometimes it is heard that 'nobody understands quantum theory.' This is wrong. In fact it is worse than wrong. It is disinformation, a habit found only in dictatorships; it is used there to prevent people from making up their own mind and from enjoying life. Quantum theory is the set of consequences that follows from the existence of a minimal action. These consequences can be understood and enjoyed by everybody. In order to do so, our first task on our way towards the top of Motion Mountain will be the study of our classical standard of motion: the motion of light.

> Nie und nirgends hat es Materie ohne Bewegung gegeben, oder kann es sie geben.
> Friedrich Engels, *Anti–Dühring*.*

Gedanken experiments and challenges

Even if we assume that experiments so far do not contradict the minimum action, we still have to check that the minimum action does not contradict reason. In particular, the minimum action must also appear in all *imagined* experiments. This is not evident.

- Angular momentum has the same unit as action. A smallest action implies that there is a smallest angular momentum in nature. How can this be, given that some particles have spin zero, i.e., have no angular momentum? Challenge 1142 n
- Could we have started the whole discussion of quantum theory by stating that there is a minimum angular momentum instead of a minimum action? Challenge 1143 n
- When electromagnetic fields play a role, the value of the action (usually) depends on the choice of the vector potential, and thus on the gauge choice. We found out in the section of electrodynamics that a suitable gauge choice can change the action value by adding or subtracting any desired amount. Nevertheless, there is a smallest action in

Ref. 654 * 'Never and nowhere has matter existed, or can it exist without motion.' Friedrich Engels (1820–1895) was one of the theoreticians of Marxism, often also called Communism.

nature. This is possible, because in quantum theory, physical gauge changes cannot add or subtract any amount, but only multiples of *twice* the minimum value. The addition property thus does not help to go below the minimum action.

▪ (Adult) plants stop to grow in the dark. Plant needs light to grow. Without light, the reactions necessary for growth stop. Can you deduce that this is a quantum effect, not explainable by classical physics?

Challenge 1144 ny

In summary, all imagined situations lead to the result that nature shows a minimum action.

20. Light – the strange consequences of the quantum of action

Alle Wesen leben vom Lichte,
jedes glückliche Geschöpfe.
Friedrich Schiller, *Wilhelm Tell*.*

What is colour?

If all the colours of materials are quantum effects, as just argued, it becomes even more interesting to study the properties of light in the light of the quantum of action. If in nature there is a minimum change, there should also be a minimum illumination. This had been already predicted by Epicurus (341–271 BCE) in ancient Greece. He stated that light is a stream of little particles, so that the smallest illumination would be that of a single light particle.

Ref. 656

In the 1930s, Brumberg and Vavilov found a beautiful way to check this prediction using the naked eye, despite our inability to detect single photons. In fact, the experiment is so simple that it could have been performed at least a century before that; but nobody had a sufficiently daring imagination to try it. The two researchers constructed a small shutter that could be opened for time intervals of 0.1 s. From the other side, in a completely dark room, they illuminated the opening with extremely weak green light of 20 aW (at 505 nm). This means that when the shutter opens, on average about 50 photons can pass, which is just the sensitivity threshold of the eye. They simply looked into the shutter repeatedly. The result is surprising but simple. Sometimes one observes light, sometimes one does not. The result is completely random. Brumberg and Vavilov gave the simple explanation: due to fluctuations, half of the time the number of photons is above eye threshold, half of the time below. The fluctuations are random, and thus the detection is as well. This would *not* happen if light would be a continuous stream; in that case, the eye would detect light at every opening of the shutter. (At higher light intensities, the percentage of non-observations quickly decreases, in accordance with the explanation.) Nobody knows what would have happened to the description of light if this simple experiment had been performed 100 years earlier.

Ref. 655

* 'All beings live of light, every happy creature.' Friedrich Schiller (b. 1759 Marbach, d. 1805 Weimar), important German poet, playwright and historian.

The experiment becomes clearer when we use devices to help us. A simple way is to start with a screen behind a prism illuminated with white light. The light is split into colours. When the screen is put further and further away, the illumination intensity cannot become infinitely small, as that would contradict the quantum of action. To check this prediction, we only need some black and white photographic film. Everybody knows that film is blackened by daylight of any colour; at medium light intensities it becomes dark grey and at lower intensities light grey. Looking at an extremely light grey film under the microscope, we discover that even under uniform illumination the grey shade actually is a more or less dense collections of black spots. Exposed film does *not* show a homogeneous colour; on the contrary, it reacts as if light is made of small particles.

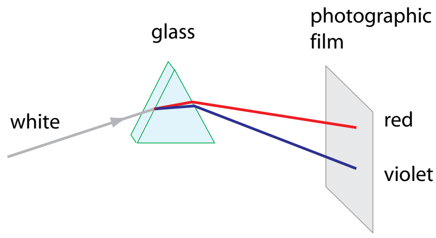


Figure 280 Illumination by pure-colour light

This is a general observation: whenever sensitive light detectors are constructed with the aim to 'see' as accurately as possible, thus in environments as dark as possible, one always finds that light manifests itself as a stream of *light quanta*. Nowadays they are usually called *photons*, a term that appeared in 1926. A low or high light intensity is simply a small or high number of photons.

Another weak source of light are single atoms. Atoms are tiny spheres; when they radiate light or X-rays, the radiation should be emitted as spherical waves. But in all experiments it is found that each atom emits only one 'blob' of light. One never finds that the light emitted by atoms forms a spherical wave, as is suggested by everyday physics. If a radiation emitting atom is surrounded by detectors, there is always only a *single* detector that is triggered.

Figure to be inserted

Figure 281 Observation of photons

Experiments in dim light thus show that the continuum description of light is *not correct*. More precise measurements confirm the role of the quantum of action: every photon leads to the *same* amount of change in the film. This amount of change is the *minimal* amount of change that light can produce. Indeed, if a minimum action would not exist, light could be packaged into arbitrary small amounts. In other words, the classical description of light by a *continuous* state function $A(t,x)$ or $F(t,x)$, whose evolution is described by a principle of least action, is *wrong*, as it does *not* describe the observed particle effects. Another, modified description is required. The modification has to be important only at low light intensities, since at high intensities the classical Lagrangian accurately describes all experimental observations.*

Page 510

At which intensities does light cease to behave as a continuous wave? Our eye can help us to find a limit. Human eyesight does not allow to consciously distinguish single

* This transition from the classical case to the quantum case used to be called *quantization*. The concept and the ideas behind it are only of historical interest today.

Ref. 657 photons, even though experiments show that the hardware of the eye is able to do this. The faintest stars which can be seen at night produce a light intensity of about 0.6 nW/m^2. Since the pupil of the eye is quite small, and as we are not able to see individual photons, photons must have energies smaller than 10^{-16} J.

In today's laboratory experiments, recording and counting individual photons is standard practice. Photon counters are part of many spectroscopy set-ups, such as those used to measure smallest concentration of materials. For example, they help to detect drugs in human hair. All these experiments thus prove directly that light is a stream of particles, as Epicurus had advanced in ancient Greece.

This and many other experiments show that a beam of light of frequency f or angular frequency ω, which determines its colour, is accurately described as a stream of photons, each with the *same* energy E given by

$$E = \hbar\, 2\pi f = \hbar\, \omega\ . \tag{476}$$

This shows that for light, the smallest measurable action is given by the quantum of action $\hbar$. This is *twice* the smallest action observable in nature; the reasons and implications will unfold during the rest of our walk. In summary, *colour* is a property of photons. A coloured light beam is a hailstorm of the corresponding photons.

The value of *Planck's constant* can be determined from measurements of black bodies Page 565 or other light sources. The result

$$\hbar = 1.1 \cdot 10^{-34}\,\mathrm{Js} \tag{477}$$

is so small that we understand why photons go unnoticed by humans. Indeed, in normal light conditions the photon numbers are so high that the continuum approximation for the electromagnetic field is of high accuracy. In the dark, the insensitivity of the signal processing of the human eye, in particular the slowness of the light receptors, makes photon Ref. 657 counting impossible. The eye is not far from maximum possible sensitivity though. From the numbers given above about dim stars we can estimate that humans are able to see Challenge 1145 ny *consciously* flashes of about half a dozen detected photons.

In the following, we will systematically deduce the remaining properties of photons, using the data collected in classical physics, while taking the quantum of action firmly into account. For example, photons have no (rest) mass and no electric charge. Can you Challenge 1146 ny confirm this? In fact, experiments can only give an upper limit for both quantities. The Ref. 658 present experimental upper limit for the (rest) mass of a photon is 10^{-51} kg.

We know that light can *hit* objects. Since the energy and the speed of photons is known, Challenge 1147 ny we guess that the photon momentum obeys

$$p = \frac{E}{c} = \hbar\,\frac{2\pi}{\lambda} \quad \text{or} \quad \mathbf{p} = \hbar\,\mathbf{k}\ . \tag{478}$$

In other words, if light is made of particles, we should be able to play billiard with them. Ref. 659 This is indeed possible, as Arthur Compton showed in a famous experiment in 1923. He directed X-rays, which are high energy photons, onto graphite, a material in which electrons move almost freely. He found that whenever the electrons in the material get hit by the X-ray photons, the deflected X-rays change colour. As expected, the strength of the hit depends on the deflection angle of the photon. From the colour change and the reflection angle, Compton confirmed that the photon momentum indeed obeys the above expression. All other experiments agree that photons have momentum. For example, when an atom *emits* light, the atom feels a *recoil*; the momentum again turns out to be given by the same value (478). In short, every photon has momentum.

The value of photon momentum respects the indeterminacy principle. In the same way that it is impossible to measure exactly both the wavelength of a wave and the position of its crest, it becomes impossible to measure both the momentum and the position of a photon. Can you confirm this? In other words, the value of the photon momentum is a direct consequence of the quantum of action.

Challenge 1148 ny

From our study of classical physics we know that light has a property beyond its colour: light can be *polarized*. That is only a complicated way to say that light can *turn* objects it shines on. In other words, light has an angular momentum oriented along the axis of propagation. What about photons? Measurements consistently find that each light particle carries an *angular momentum* given by $L = \hbar$. It is called its *helicity*; to make it more clear that the quantity is similar to that found for massive particles, one also speaks of the *spin* of photons. Photons somehow 'turn' – in a sense either parallel or antiparallel to the direction of motion. Again, the magnitude of the photon helicity or spin is not a surprise; it confirms the classical relation $L = E/\omega$ between energy and angular momentum that we found in the section on classical electrodynamics. Note that in contrast to intuition, the angular momentum of a single photon is fixed, and thus *independent* of its energy. Even the photons with the highest energy have $L = \hbar$. Of course, the value of the helicity also respects the limit given by the quantum of action. The helicity value $\hbar$ – twice the minimum $\hbar/2$ – has important consequences which will become clear shortly.

Page 530

Challenge 1149 ny

What is light? – Again

> La lumière est un mouvement luminaire de corps lumineux.*
>
> Blaise Pascal

In the seventeenth century, Blaise Pascal** used this sentence to make fun about certain physicists. He ridiculed (rightly so) the blatant use of a circular definition. Of course, he was right; in his time, the definition was indeed circular, as no meaning could be given to any of the terms. But as usual, whenever an observation is studied with care by physicists, they give philosophers a beating. All those originally undefined terms now have a definite meaning. Light is indeed a type of motion; this motion can rightly be called luminary because in opposition to motion of material bodies, it has the unique property $v = c$; the luminous bodies, today called photons, are characterized and differentiated from all other particles by their dispersion relation $E = cp$, their energy $E = \hbar\omega$, their spin $L = \hbar$, the vanishing value of all other quantum numbers, and by being the quanta of the electromagnetic field.

* Light is the luminary movement of luminous bodies.

** Blaise Pascal (b. 1623 Clermont, d. 1662 Paris) important French mathematician and physicist up to the age of twenty-six; he then turned theologian and philosopher.

In short, *light is a stream of photons*. It is indeed a luminary movement of luminous bodies. The existence of photons is the first example of a general property of the world on small scales: *all waves and all flows in nature are made of quantum particles*. Large numbers of (coherent) quantum particles – or quantons – do indeed behave as waves. We will see shortly that this is the case even for matter. The fundamental constituents of *all* waves are quantons. There is *no* exception. The everyday, continuum description of light is thus similar in many aspects to the description of water as a continuous fluid; photons are the atoms of light, and continuity is an approximation valid for large particle numbers. Small numbers of quantons often behave like classical particles. In the old days of physics, books used to discuss at length a so-called *wave–particle duality*. Let us be clear from the start: quantons, or quantum particles, are *neither* classical waves *nor* classical particles. In the microscopic world, quantons are the fundamental objects.

However, a lot is still unclear. Where inside matter do these monochromatic photons come from? Even more interestingly, if light is made of quantons, all electromagnetic fields, even static ones, must be made of photons as well. However, in static fields nothing is flowing. How is this apparent contradiction solved? And what effects does the particle aspect have on these static fields? What is the difference between quantons and classical particles? The properties of photons thus require some more careful study. Let us go on.

The size of photons

First of all, we might ask: what are these photons made of? All experiments so far, performed down to the present limit of about 10^{-20} m, give the same answer: 'we can't find anything'. That is consistent both with a vanishing mass and a vanishing size of photons; indeed, one intuitively expects any body with a finite size to have a finite mass. Thus, even though experiments give only an upper limit, it is consistent to claim that a photon has *no size*.

A particle with no size cannot have any constituents. A photon thus cannot be divided into smaller entities. For this reason people refer to photons as *elementary* particles. We will give some strong additional arguments for this deduction soon. (Can you find one?)

Challenge 1150 ny

This is a strange result. How can a photon have vanishing size, have no constituents, and still be *something*? This is a hard questions; the answer will appear only later on. At the moment we simply have to accept the situation as it is. We therefore turn to an easier issue.

Are photons countable? – Squeezed light

Also gibt es sie doch.

Max Planck*

Above we saw that in order to count photons, the simplest way is to distribute them across a large screen and then to absorb them. But this method is not fully satisfying, as it destroys the photons. How can one count photons without destroying them?

* 'Thus they do exist after all.' Max Planck, in later years, said this after standing silently, for a long time, in front of an apparatus which counted single photons by producing a click for each photon it detected. It is not a secret that for a large part of his life, Planck was not a friend of the photon concept, even though his own results were the starting point for its introduction.

One way is to reflect photons on a mirror and to measure the recoil of the mirror. This seems almost unbelievable, but nowadays the effect is becoming measurable even for small number of photons. For example, this effect has to be taken into account in the mirrors used in gravitational wave detectors, where the position of laser mirrors has to be measured to high precision.

Another way of counting photons without destroying them uses special high quality laser cavities. Using smartly placed atoms inside such a cavity, it is possible to count the number of photons by the effect they have on these atoms.

In other words, light intensity can indeed be measured without absorption. However, the next difficulty appears straight away. Measurements show that even the best light beams, from the most sophisticated lasers, *fluctuate* in intensity. There are no steady beams. This does not come as a surprise: if a light beam would *not* fluctuate, observing it twice in a row would yield a vanishing value for the action. However, there is a minimum action in nature, namely $\hbar/2$. Thus any beam and any flow in nature fluctuates. But there is more.

A light beam is described by its intensity and its phase. The change – or action – occurring while a beam moves is given by the variation in the product of intensity and phase. Experiments confirm the obvious deduction: intensity and phase of beams behave like momentum and position of particles: they obey an indeterminacy relation. You can deduce it yourself, in the same way we deduced Heisenberg's relations. Using as characteristic intensity $I = E/\omega$ the energy per circular frequency and calling the phase φ, we get*

$$\Delta I \, \Delta\varphi \geqslant \frac{\hbar}{2} \; . \qquad (480)$$

For light emitted from usual lamps, the product of the left side is much larger than the quantum of action. On the other hand, laser beams can (almost) reach the limit. Among these, light beams in which the two indeterminacies strongly differ from each other are called *non-classical light* or *squeezed light*; they are used in many modern research applications. Such light beams have to be treated carefully, as the smallest disturbances transform them back into usual laser beams, where the two indeterminacies have the same value. An example of non-classical light are those beams with a given, fixed photon number, thus with an extremely large phase indeterminacy.

The observation of non-classical light points to a strange consequence valid even for classical light: the number of photons in a light beam is *not* a defined quantity. In general it is *undetermined*, and it *fluctuates*. The number of photons at the beginning of a beam is not necessarily the same as at the end of the beam. Photons, in contrast to stones, cannot be counted precisely – as long as they move. It is only possible to detect an approximate number, within the limit set by indeterminacy.

* A large photon number is assumed in the expression; this is obvious, as $\Delta\varphi$ cannot grow beyond all bounds. The exact relations are

$$\Delta I \, \Delta \cos\varphi \geqslant \frac{1}{2} |\langle \sin\varphi \rangle|$$
$$\Delta I \, \Delta \sin\varphi \geqslant \frac{1}{2} |\langle \cos\varphi \rangle| \qquad (479)$$

where $\langle x \rangle$ denotes the expectation value of the observable x.

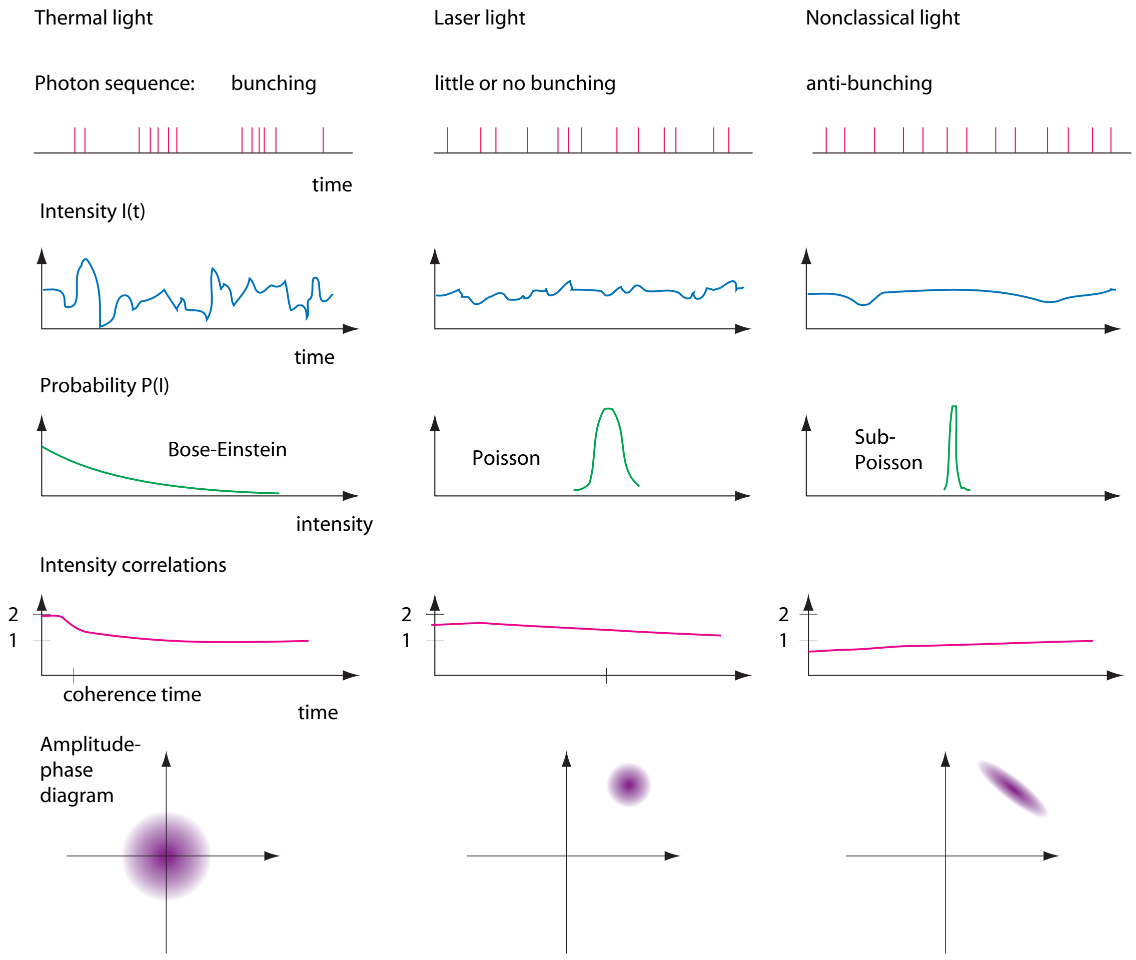


Figure 282 Various types of light

The most extreme example are those light beams with an (almost) fixed phase. In such beams, the photon number fluctuates from zero to infinity. In other words, to produce a *coherent* laser beam, approximating a pure sine wave as perfectly as possible, one must build a source in which the photon number is as undetermined as possible.

The other extreme is a beam with a fixed number of photons; in such a beam, the phase fluctuates erratically. Most daily life situations, such as the light from incandescent lamps, lie somewhere in the middle: both phase and intensity indeterminacies are of similar magnitude.

As an aside, it turns out that in deep, dark intergalactic space, far from every star, there still are about 400 photons per cubic centimetre. But also this number, like the number of photons in a light beam, has its measurement indeterminacy. Can you estimate it?

Challenge 1151 ny

In summary, unlike little stones, *photons are not countable*. And this it not the last difference between photons and stones.

The position of photons

Where is a photon when it moves in a beam of light? Quantum theory gives a simple answer: nowhere in particular. The proof is given most spectacularly by experiments with interferometers; they show that even a beam made of a *single* photon can be split, be led along two different paths and then be recombined. The resulting interference shows that the single photon cannot be said to have taken either of the two paths. If one of the two paths is blocked, the pattern on the screen changes. In other words, the photons somehow must have taken both paths at the same time. Photons cannot be localized; they have no position.*

This impossibility of localization can be specified. It is impossible to localize photons in the direction *transverse* to the motion. It is less difficult to localize photons *along* the motion direction. In the latter case, the quantum of action implies that the longitudinal position is uncertain within a value given by the wavelength of the corresponding colour.

Challenge 1152 ny

Can you confirm this?

In particular, this means that photons *cannot* be simply visualized as short wave trains. Photons are truly *unlocalizable* entities specific to the quantum world.

Now, if photons can almost be localized along their motion, we can ask the following question: How are photons lined up in a light beam? Of course, we just saw that it does not make sense to speak of their *precise* position. But are photons in a perfect beam arriving in almost regular intervals or not?

Ref. 660

To the shame of physicists, the study of this question was started by two astronomers, Robert Hanbury Brown and Richard Twiss, in 1956. They used a simple method to measure the probability that a second photon in a light beam arrives at a given time after a first one. They simply split the beam, put one detector in the first branch and varied the position of a second detector in the other branch.

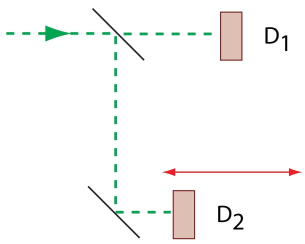


Figure 283 How to measure photon statistics

Hanbury Brown and Twiss found that for coherent light the clicks in the two counters, and thus the photons, are *correlated*. This result is in complete contrast with classical electrodynamics. Photons are indeed necessary to describe light. In more detail, their experiment showed that whenever a photon hits, the probability that a second one hits just afterwards is highest. Photons in beams are thus *bunched*.

Every light beam shows an upper limit time to bunching, called the *coherence time*. For times larger than the coherence time, the probability for bunching is low and independent of the time interval, as shown in Figure 283. The coherence time characterizes every light beam, or better, every light source. In fact, it is more intuitive to use the concept of *coherence length*, as it gives a clearer image for a light beam. For thermal lamps, the coherence length is only a few micrometers, a small multiple of the wavelength. The largest coherence lengths, up to over 100 000 km, are realized in research lasers. Interestingly, coherent light is even found in nature; several special stars have been found to emit coherent light.

Ref. 661

* This conclusion cannot be avoided by saying that photons are *split* at the beam splitter: if one puts a detector into each arm, one finds that they never detect a photon at the same time. Photons cannot be divided.

Even though the intensity of a good laser light is almost constant, laser beam photons still do not arrive in regular intervals. Even the best laser light shows bunching, though with different statistics and to a lower degree than lamp light. Light for which photons arrive regularly, thus showing so-called *(photon) anti-bunching*, is obviously non-classical in the sense defined above; such light can be produced only by special experimental arrangements. The most extreme example is pursued at present by several research groups; they aim to construct light sources which emit one photon at a time, at regular time intervals, as reliably as possible.

In summary, experiments force us to conclude that light is made of photons, but that photons *cannot* be localized in light beams. It makes no sense to talk about the position of a photon in general; the idea makes only sense in some special situations, and then only approximately and as a statistical average.

Are photons necessary?

In light of the results uncovered so far, the answer to the title question is obvious. But the issue is tricky. In school books, the *photoelectric effect* is usually cited as the first and most obvious experimental proof of the existence of photons. In 1887, Heinrich Hertz observed that for certain metals, such as lithium or caesium, incident ultraviolet light leads to charging of the metal. Later it was shown that the light leads to the emission of electrons, and that that the energy of the ejected electrons is *not* dependent on the intensity of the light, but only dependent on the difference between $\hbar$ times its frequency and a material dependent threshold energy. In 1905, Albert Einstein predicted this result from the assumption that light is made of photons of energy $E = \hbar\omega$. He imagined that this energy is used partly to extract the electron over the threshold, and partly to give it kinetic energy. More photons only lead to more electrons, not to faster ones.

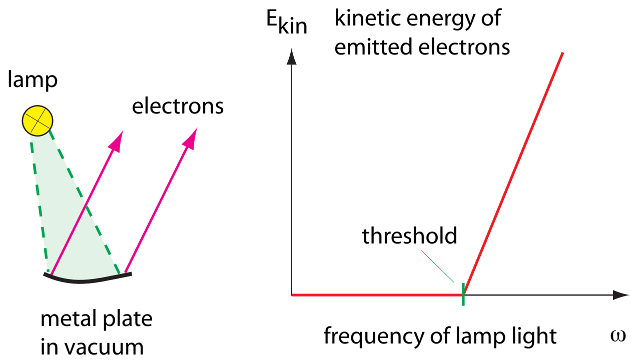


Figure 284 The kinetic energy of electrons emitted in the photoelectric effect

Ref. 663

Einstein received the Nobel price for this explanation. But Einstein was a *genius*; that means he deduced the correct result by a somewhat incorrect reasoning. The (small) mistake was the prejudice that a classical, continuous light beam would produce a different effect. It does not take a lot to imagine that a classical, continuous electromagnetic field interacting with discrete matter, made of discrete atoms containing discrete electrons, leads to exactly the same result, if the motion of electrons is described by quantum theory. Several researchers confirmed this point already early in the twentieth century. The photoelectric effect by itself does *not* require photons for its explanation.

Ref. 664

Many researchers were unconvinced by the photoelectric effect. Historically, the most important argument for the *necessity* of light quanta was given by Henri Poincaré. In 1911 and 1912, at age 57 and only a few months before his death, he published two influential papers proving that the radiation law of black bodies, the one in which the quantum Ref. 665 of action had been discovered by Max Planck, *requires* the introduction of photons. He also showed that the amount of radiation emitted by a hot body is *finite* only due to the *quantum* nature of the processes leading to light emission. A description of the processes by classical electrodynamics would lead to *infinite* amounts of radiated energy. These two influential papers convinced most of the sceptic physics researchers at the time that it was worthwhile to study quantum phenomena in more detail. Poincaré did not know about the action limit $S \geqslant \hbar/2$; yet his argument is based on the observation that light of a given frequency always has a minimum intensity, namely one photon. Splitting such a one photon beam into two beams, e.g. using a half-silvered mirror, does produce two beams. However, there is no way to find more than one photon in those two beams together.

Another interesting experiment requiring photons is the observation of 'molecules of Ref. 666 photons'. In 1995, Jacobson et al. predicted that the de Broglie wavelength of a *packet* of photons could be observed. Following quantum theory it is given by the wavelength of a single photon divided by the number of photons in the packet. The team argued that the packet wavelength could be observable if one would be able to split and recombine such packets without destroying the cohesion within the packet. In 1999, this effect was indeed observed by de Pádua and his brazilian research group. They used a careful set-up with a nonlinear crystal to create what they call a *biphoton*, and observed its interference properties, finding a reduction of the effective wavelength by the predicted factor of two. In the meantime, packages with three and even four entangled photons have been created Ref. 667 and observed.

Still another argument for the necessity of photons is the mentioned recoil felt by atoms emitting light. The recoil measured in these cases is best explained by the emission of a photon in a particular direction. In contrast, classical electrodynamics predicts the emission of a spherical wave, with no preferred direction.

Obviously, the observation of *non-classical light*, also called *squeezed light*, also argues for the existence of photons, as squeezed light proves that photons indeed are an *intrinsic* Ref. 668 aspect of light, necessary even when no interactions with matter play a role. The same is true for the Hanbury Brown–Twiss effect.

Finally, the spontaneous decay of excited atomic states also requires the existence of photons. A continuum description of light does not explain the observation.

In summary, the concept of photon is indeed *necessary* for a precise description of light, but the details are often subtle, as the properties of photons are unusual and require a change in thinking habits. To avoid these issues, all school books stop discussing photons after the photoelectric effect. That is a pity; things get interesting only after that. To savour the fascination, ponder the following issue. Obviously, all electromagnetic fields are made of photons. Photons can be counted for gamma rays, X-rays, ultraviolet light, visible light and infrared light. However, for lower frequencies, photons have not been detected yet. Can you imagine what would be necessary to count the photons emitted from a radio Challenge 1153 ny station?

The issue directly leads to the most important question of all:

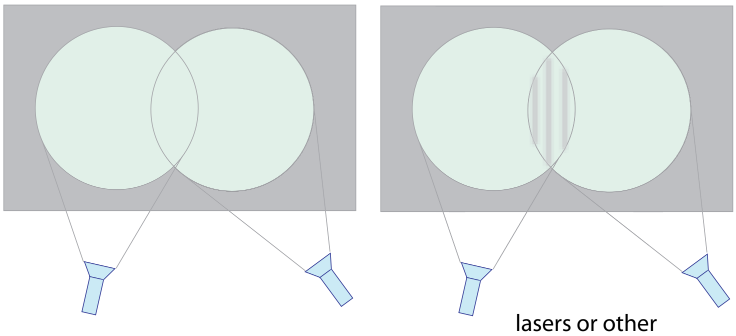


Figure 285 Light crossing light

How can a wave be made up of particles?

> Fünfzig Jahre intensiven Nachdenkens haben mich der Antwort auf die Frage 'Was sind Lichtquanten?' nicht näher gebracht. Natürlich bildet sich heute jeder Wicht ein, er wisse die Antwort. Doch da täuscht er sich.
>
> Albert Einstein, 1951 *

If a light wave is made of particles, one must be able to explain each and every wave properties with the help of photons. The experiments mentioned above already hinted that this is possible only because photons are *quantum* particles. Let us take a more detailed look at this connection.

Light can *cross* other light undisturbed. This observation is not hard to explain with photons; since photons do not interact with each other, and since they are point-like, they 'never' hit each other. In fact, there indeed *is* an extremely small probability for their interaction, as will be found below, but this effect is not observable in everyday life.

But the problems are not finished yet. If two light beams of identical frequencies and fixed phase relation cross, we observe alternating bright and dark regions, so-called *interference fringes*. How do these interference fringes appear? Obviously, photons are not detected in the dark regions. How can this be? There is only one possible way to answer: the brightness gives the *probability* for a photon to arrive at that place. The fringes imply that photons behave like little *arrows*. Some additional thinking leads to the following description:

(1) the probability of a photon arriving somewhere is given by the *square of an arrow*;
(2) the final arrow is the *sum* of all arrows getting there, taking all possible paths;
(3) the arrow's direction stays fixed in space when photons move;
(4) the length of an arrow shrinks with the square of the travelled distance;

* 'Fifty years of intense reflection have not brought me nearer to the answer of the question 'What are light quanta?' Of course nowadays every little mind thinks he knows the answer. But he is wrong.'

(5) photons emitted by one-coloured sources are emitted with arrows of constant length pointing in direction ωt; in other words, such monochromatic sources spit out photons with a *rotating* mouth.

(6) photons emitted by thermal sources, such as pocket lamps, are emitted with arrows of constant length pointing in *random* directions.

With this model* we can explain the stripes seen in laser experiments, such as those of Figure 285 and Figure 286. You can check that in some regions, the two arrows travelling through the two slits add up to zero *for all times*. No photons are detected there. In other regions, the arrows always add up to the maximal value. These regions are always bright. In between regions give in between shades. Obviously, for the case of pocket lamps the brightness also behaves as expected: the averages then simply add up, as in the common region in the left case of Figure 285.

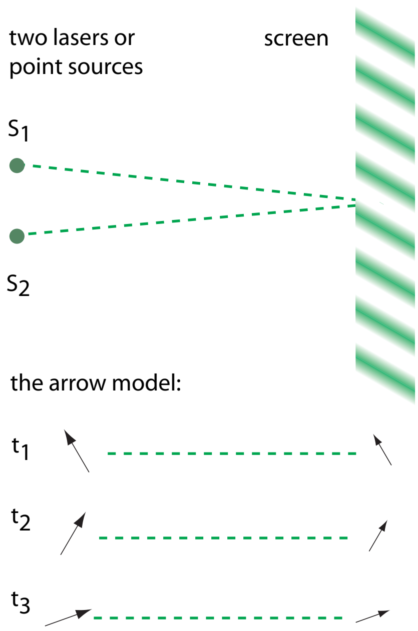


Figure 286 Interference and the description of light with arrows (at one particular instant of time)

You might want to calculate the distance of the lines when the source distance, the colour and the distance to the screen is given. Challenge 1154 ny

Obviously, the photon model implies that interference patterns are built up as the sum of a large number of one-photon hits. Using low intensity beams, we should therefore be able to see how these little spots slowly build up an interference pattern by accumulating at the bright spots and never hitting the dark regions. That is indeed the case. All experiments have confirmed this description.

It is important to stress that interference of two light beams is *not* the result of two different photons cancelling out or adding each other up. The cancelling would contradict energy and momentum conservation. Interference is an effect valid for each photon separately, because each photon is spread out over the whole set-up; each photon takes all possible paths and interferes. As Paul Dirac said, *each photon interferes only with itself.* Ref. 669 Interference only works because photons are quantons, and not at all classical particles.

Dirac's widely cited statement leads to a famous paradox: if a photon can interfere only with itself, how can two laser beams from two different lasers show interference? The answer of quantum physics is simple but strange: in the region where the beams interfere, there is no way to say from which source a photon is arriving. Photons are quantons; the photons in the crossing region *cannot* be said to come from a specific source. Photons in the interference region are quantons on their own right, which indeed interfere only with themselves. In that region, one cannot honestly say that light is a flow of photons. Despite regular claims of the contrary, Dirac's statement is correct. That is the strange result of the quantum of action.

Waves also show *diffraction*. To understand this phenomenon with photons, let us

* The model gives a correct description of light with the exception that it neglects polarization.

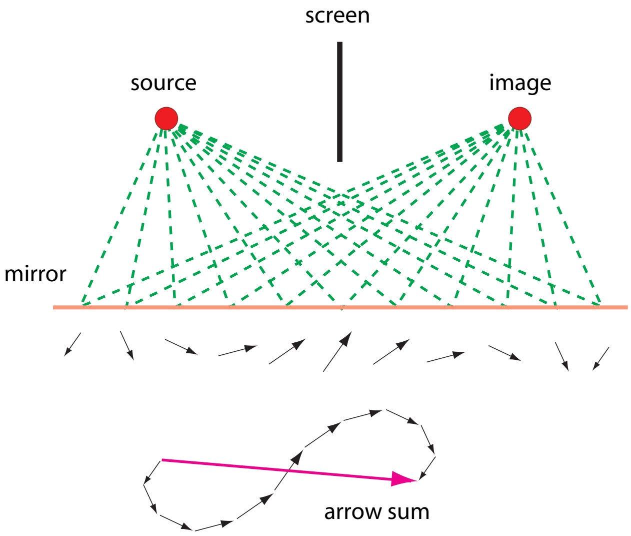


Figure 287 Light reflected by a mirror and the corresponding arrows (at one particular instant of time)

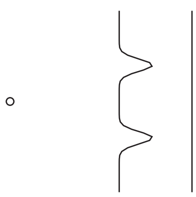
start with a simple mirror and study *reflection* first. Photons (like any quantum particle) move from source to detector in *all* ways possible. As the discoverer of this explanation, Richard Feynman,* likes to stress, the term 'all' has to be taken literally. This was not a big deal in the explanation of interference. But in order to understand a mirror we have to include all possibilities, as crazy as they seem, as shown in Figure 287.

For a mirror, we have to add up the arrows arriving at the same time at the location of the image. They are shown, for each path, below the corresponding segment of the mirror. The arrow sum shows that light indeed does arrive at the image. It also shows that most of the contributions is coming from those paths near the middle one. If we were to perform the same calculation for another direction, (almost) no light would get there. In summary, the rule that reflection occurs with incoming angle equal to the outgoing angle is an approximation only; the rule follows from the arrow model of light.

In fact, a detailed calculation, with more arrows, shows that the approximation is quite precise; the errors are much smaller than the wavelength of the light used.

The proof that light does indeed take all these strange paths is given by a more specialized mirror. As show in Figure 288, one can repeat the experiment with a mirror which reflects only along certain stripes. In this case, the stripes were carefully chosen such that

* Richard ('Dick') Phillips Feynman (b. 1918 New York City, d. 1988), US American physicist. One of the founders of quantum electrodynamics, he discovered the 'sum-over-histories' reformulation of quantum theory, made important contributions to the theory of the weak interaction and of quantum gravity, and co-authored a famous physics textbook, the *Feynman Lectures on Physics*. He is one of those theoretical physicists who made career mainly by performing complex calculations, a fact he tried to counter at the end of his life. Though he tried to surpass the genius of Wolfgang Pauli throughout his whole life, he failed in this endeavour. He was famously arrogant, disrespectful of authorities, as well as deeply dedicated to physics and to enlarging knowledge in his domain. He also was a well known collector of surprising physical explanations and an author of several popular texts on his work and his life. He shared the 1965 Nobel prize in physics for his work on quantum electrodynamics.



the arrows reflected there all show a bias to one direction, namely to the left. The same calculation now shows that such a specialized mirror, usually called a *grating*, allows light to be reflected in unusual directions. And indeed, this behaviour is standard for waves and is called *diffraction*. In short, the arrow model for photons does allow to describe this wave property of light, provided that photons follow the mentioned crazy probability scheme. Do not get upset; as said before, quantum theory *is* the theory of crazy people.

You may want to check that the arrow model, with the approximations it generates by summing over all possible paths, automatically ensures that the quantum of action is indeed the smallest action that can
Challenge 1155 ny be observed.

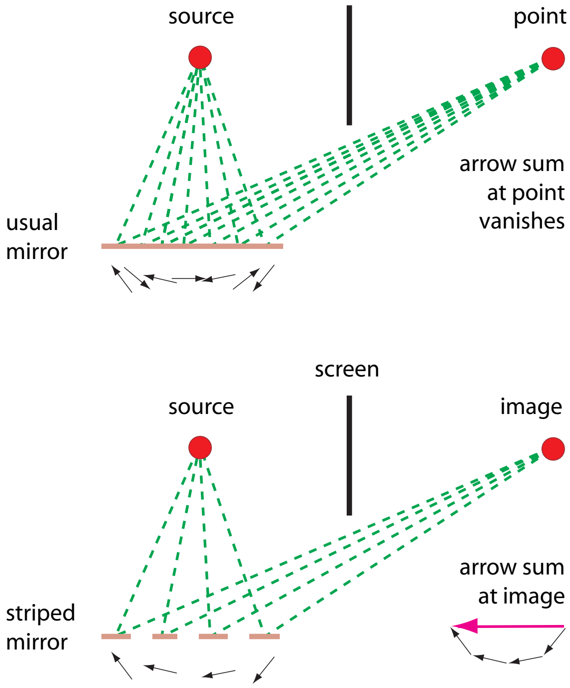


Figure 288 The light reflected by a badly placed mirror and by a grating

All waves have a *signal velocity*. As a consequence, waves show *refraction* when they move from one medium into another with different signal velocity. Interestingly, the naive particle picture of photons as little stones would imply that light is faster in materials with high indices of refraction, the so-called *dense* materi-
Challenge 1156 ny als. Just try it. However, experiments show that light in dense materials moves *slowly*. The wave picture has no difficulties explaining this obser-
Challenge 1157 ny vation. (Can you confirm it?) Historically, this was one of the arguments *against* the particle theory of light. However, the arrow model of light
Challenge 1158 ny presented above is able to explain refraction properly. It is not difficult doing so; try it.

Waves also *reflect partially* from materials such as glass. This is one of the toughest properties of waves to be explained with photons. The issue is important, as it is one of the few effects that is *not* explained by a classical wave theory of light. However, it *is* explained by the arrow model, as we will find out shortly. Partial reflection confirms the description of the rules (1) and (2) of the arrow model. Partial reflection shows that photons indeed behave *randomly:* some are reflected and other are not, without any selection criterion. The distinction is purely statistical. More about this issue shortly.

In waves, the fields *oscillate in time and space*. One way to show how waves can be made of particles is to show once for all how to build up a sine wave using a large number of photons. A sine wave is a coherent state of light. The way to build them up was
Ref. 670 explained by Glauber. In fact, to build a pure sine wave, one needs a superposition of a beam with one photon, a beam with two photons, a beam with three photons, continuing up to a beam with an infinite number of them. Together, they give a perfect sine wave. As expected, its photon number fluctuates to the highest degree possible.

If we repeat the calculation for non-ideal beams, we find that the indeterminacy relation for energy and time is respected; every emitted wave will possess a certain spectral width. Purely monochromatic light does not exist. Similarly, no system which emits a wave *at random* can produce a monochromatic wave. All experiments confirm these results.

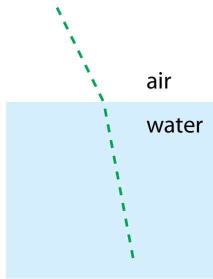


Figure 289 If light were made of little stones, they would move faster inside water

Waves can be *polarized*. So far, we disregarded this property. In the photon picture, polarization is the result of carefully superposing beams of photons spinning clockwise and anticlockwise. Indeed, we know that linear polarization can be seen as a result of superposing circularly polarized light of both signs, using the proper phase. What seemed a curiosity in classical optics turns out to be the fundamental explanation of quantum theory.

Photons are *indistinguishable*. When two photons of the same colour cross, there is no way to say, after the crossing, which of the two is which. The quantum of action makes this impossible. The indistinguishability of photons has an interesting consequence. It is impossible to say which emitted photon corresponds to which arriving photon. In other words, there is *no* way to follow the path of a photon in the way we are used to follow the path of a billiard ball. Particles which behave in this way are called bosons. We will discover more details about the indistinguishability of photons in the next chapter.

In summary, we find that light waves *can* indeed be built of particles. However, this is only possible under the condition that photons are not precisely countable, that they are not localizable, that they have no size, no charge and no mass, that they carry an (approximate) phase, that they carry spin, that they are indistinguishable bosons, that they can take any path whatsoever, that one cannot pinpoint their origin and that their probability to arrive somewhere is determined by the square of the sum of amplitudes for all possible paths. In other words, light can be made of particles only under the condition that these particles have extremely special, *quantum* properties. Only these quantum properties allow them to behave like waves, in the case that they are present in large numbers.

Quantons are thus quite different from usual particles. In fact, one can argue that the only (classical) particle aspects of photons are their quantized energy, momentum and spin. In all other aspects photons are *not* like little stones. It is more honest to say that Ref. 671 *photons are calculating devices to precisely describe observations about light*. Often these calculating devices are called *quantons*. In summary, *all waves are streams of quantons*. In fact, all waves are *correlated* streams of quantons. That is true both for light, for any other form of radiation, as well as for matter, in all its forms.

The strange properties of quantons are the reason that earlier attempts to describe light as a stream of (classical) particles, such as the one by Newton, failed miserably, under the rightly deserved ridicule of all other scientists. Indeed, Newton upheld his idea against all experimental evidence, especially that on light's wave properties, something a physicist should never do. Only when people accepted that light is a wave and then discovered and understood that quantum particles are different from classical particles was the approach successful.

The indeterminacy relations show that even a single quanton can be seen as a wave; however, whenever it interacts with the rest of the world, it behaves as a particle. In fact it is *essential* that all wave are made of quantons; if not, interactions would not be local, and objects, in contrast to experience, could not be localized at all. To separate between wave and particle descriptions, we can use the following criterion. Whenever matter and light interact, it is more appropriate to describe electromagnetic radiation as a wave if the wavelength λ obeys

$$\lambda \gg \frac{\hbar c}{kT} , \tag{481}$$

where $k = 1.4 \cdot 10^{-23}$ J/K is Boltzmann's constant. If the wavelength is much *smaller* than the right hand side, the particle description is most appropriate. If the two sides are of the same order of magnitude, both effects play a role.

Can light move faster than light? – Virtual photons

Light can move faster than c in vacuum, as well as slower than c. The quantum principle even explains the details. As long as the quantum principle is obeyed, the speed of a short light flash can differ a bit from the official value, though only a *tiny* bit. Can you estimate
Challenge 1159 ny the allowed difference in arrival time for a light flash from the dawn of times?

The little arrow explanation also gives the same result. If one takes into account the crazy possibility that photons can move with any speed, one finds that all speeds very different from c cancel out. The only variation that remains, translated in distances, is
Challenge 1160 ny the indeterminacy of about one wavelength in the longitudinal direction which we mentioned already above.

However, the most absurd results of the quantum of action appear when one studies *static* electric fields, such as the field around a charged metal sphere. Obviously, such a field must also be made of photons. How do they move? It turns out that static electric fields are built of *virtual* photons. In the case of static electric fields, virtual photons are *longitudinally* polarized, do *not* carry energy away, and cannot be observed as free particles. Virtual photons are photons who do not appear as free particles, but only have extremely short-lived appearances before they disappear again. In other words, photons, like any other virtual particle, are 'shadows' of particles that obey

$$\Delta x \Delta p \leqslant \hbar/2 . \tag{482}$$

Virtual particles do not obey the uncertainty relation, but its opposite; the relation expresses their short-lived appearance. Despite their intrinsically short life, they have im-
Page 709 portant effects. We will explore virtual particles in detail shortly.

In fact, the vector potential A allows *four* polarizations, corresponding to the four coordinates (t, x, y, z). For the photons one usually talks about, the free or *real* photons, the polarizations in t and z direction cancel out, so that one observes only the x and y polarizations. For bound or *virtual* photons, the situation is different.

– CS – more to be written – CS –

In short, static electric and magnetic fields are continuous flows of virtual photons. Virtual photons can have mass, can have spin directions not pointing along the motion path, and can have momentum opposite to their direction of motion. All these properties are different from real photons. In this way, exchange of virtual photons leads to the *attraction* of bodies of different charge. In fact, virtual photons *necessarily* appear in any description of electromagnetic interactions; more about their effects, such as the famous attraction of neutral bodies, will be discussed later on.

In summary, light can indeed move faster than light, though only in amounts allowed by the quantum of action. For everyday situations, i.e. for cases with a high value of the action, all quantum effects average out, including light velocities different from c.

A different topic also belongs into this section. Not only the position, but also the energy of a single photon can be undefined. For example, certain materials split one photon (Ref. 672) of energy $\hbar\omega$ into two photons, whose two energies sum up to the original one. Quantum mechanics makes the strange prediction that the precise way the energy is split is known only when the energy of one of the two photons is measured. Only at that very instant the energy of the second photon is known. Before that, both photons have undefined energies. The process of energy fixing takes place *instantaneously*, even if the second photon is far away. Page 747 We will explain below the background of this and similar strange effects, which seem to be faster than light but which are not. Indeed, such effects do not transmit energy or information faster than light. Challenge 1161 ny

Indeterminacy of electric fields

We saw that the quantum of action implies an indeterminacy for light intensity. That implies a similar limit for electric and magnetic fields. This conclusion was first drawn in 1933 by Bohr and Rosenfeld. (Ref. 673) They started from the effects of the fields on a test particle of mass m and charge q, which are described by

$$m\mathbf{a} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) . \tag{483}$$

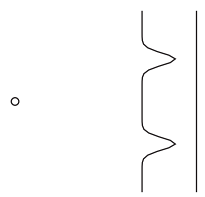
Since it is impossible to measure momentum and position of a particle, they deduced an indeterminacy for the electrical field given by (Challenge 1162 ny)

$$\Delta E = \frac{\hbar}{q\,\Delta x\,T} , \tag{484}$$

where T is the measurement time and Δx is the position uncertainty. Every value of an electric field, and similarly that of every magnetic field, is thus affected with an indeterminacy. The physical state of the electromagnetic field behaves like the state of matter in this aspect. This is the topic we explore now.

Curiosities and fun challenges about photons

- Can diffraction be explained with photons? Newton was not able to do so. Today we can do so. Figure 290 is translationally invariant along the horizontal direction; therefore, the momentum component along this direction is also conserved: $p_1 \sin\alpha_1 = p_2 \sin\alpha_2$. The photon energy $E = E_1 = E_2$ is obviously conserved.



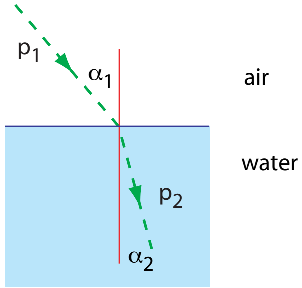


Figure 290 Diffraction and photons

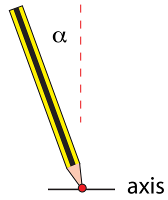


Figure 291 A falling pencil

The index of refraction n is defined with momentum and energy as

$$n = \frac{cp}{E} . \tag{485}$$

As a result, the 'law' of refraction follows.

There is an important issue here. The velocity of a photon $\mathbf{v} = \delta\mathbf{E}/\delta\mathbf{p}$ in a light ray is *not* the same as the phase velocity $\mathbf{u} = \mathbf{E}/\mathbf{p}$ that enters in the calculation.

21. Motion of matter – beyond classical physics

> All great things begin as blasphemies.
>
> George Bernard Shaw

The existence of a smallest action has numerous effects on the motion of matter. We start with a few experimental results that show that the quantum of action is indeed the smallest action.

Wine glasses and pencils

A simple consequence of the quantum of action is the impossibility of completely filling a glass of wine. If we call 'full' a glass at maximum capacity (including surface tension effects, to make the argument precise), we immediately see that the situation requires complete rest of the liquid's surface; however, the quantum of action forbids this. Indeed, a completely quiet surface would allow two subsequent observations which differ by less than $\hbar/2$. *There is no rest in nature.* In other words, the quantum of action proves the old truth that a glass of wine is always partially empty *and* partially full. Challenge 1163 n

The quantum of action has many similar consequences for everyday life. For example, a pencil on its tip *cannot* stay vertical, even if it is isolated from all disturbances, such as vibrations, air molecules and thermal motion. Are you able to confirm this? In fact, it is even possible to calculate the time after which a pencil must have fallen over.* Challenge 1164 ny

Ref. 674 * That is not easy, but neither too difficult. For an initial orientation *close* to the vertical, the fall time T

Cool gas

Rest is impossible in nature. Even at lowest temperatures, particles inside matter are in motion. This fundamental lack of rest is said to be due to the so-called *zero-point fluctuations*. A good example are the recent measurements of Bose–Einstein condensates, systems with a small number of atoms (between ten and a few million) at lowest temperatures (around 1 nK). These cool gases can be observed with high precision. Using elaborate experimental techniques, Bose–Einstein condensates can be put into states for which $\Delta p \Delta x$ is almost exactly equal to $\hbar/2$, though never lower than this value.

That leads to an interesting puzzle. In a normal object, the distance between the atoms is much larger than their de Broglie wavelength. (Are you able to confirm this?) But today it is possible to cool objects to very low temperatures. At extremely low temperatures, less than 1 nK, the wavelength of the atoms may be larger than their separation. Can you imagine what happens in such cases?

Challenge 1167 ny

Ref. 685

Challenge 1168 ny

No rest

> Otium cum dignitate.*
>
> Cicero, *De oratore.*

The impossibility of rest, like all other unexplained effects of classical physics, is most apparent in domains where the action is near the minimum observable one. To make the effects most obvious, we study the smallest amount of matter that can be isolated: a single particle. Later on we will explore situations that cover higher numbers of particles.

Experiments show that perfect rest is never observed. The quantum of action prevents this in a simple way. Whenever the position of a system is determined to high precision, we need a high energy probe. Indeed, only a high energy probe has a wavelength small enough to allow a high precision for position measurements. As a result of this high energy however, the system is disturbed. Worse, the disturbance itself is also found to be imprecisely measurable. There is thus no way to determine the original position even by taking the disturbance itself into account. In short, perfect rest cannot be observed. All systems who have ever been observed with high precision confirm that perfect rest does not exist. Among others, this result has been confirmed for electrons, neutrons, protons, ions, atoms, molecules and crystals.

---

turns out to be

$$T = \frac{1}{2\pi} T_0 \ln \frac{8}{\alpha} \qquad (486)$$

where α is the starting angle, and a fall by π is assumed. Here T_0 is the oscillation time of the pencil for small angles. (Can you determine it?)

Challenge 1165 ny

The indeterminacy relation for the tip of the pencil yields a minimum starting angle, because the momentum indeterminacy cannot be made as large as wanted. You should be able to provide an upper limit. Once the angle is known, you can calculate the maximum time.

Challenge 1166 ny

* 'Rest with dignity.'

Flows and the quantization of matter

Die Bewegung ist die Daseinsform der Materie.
Friedrich Engels, *Anti-Dühring*.*

Not only is rest made impossible by the quantum of action; the same impossibility applies to any situation which does not change in time, like any constant velocity. The most important example are *flows*. The quantum of action implies that no flow can be stationary. More precisely, a smallest action implies that all flows are made of smallest entities. All flows are made of quantum particles. Two flows ask for direct confirmation: flows of electricity and flows of liquids.

If electrical current would be a continuous flow, it would be possible to observe action values as small as desired. The simplest confirmation of the discontinuity of current flow was discovered only in the 1990s: take two metal wires on the table, laying across each other. It is not hard to let a current flow from one wire to the other, via the cross-over, and to measure the voltage. A curve like the one shown in Figure 292 is found: the current increases with voltage in regular steps.

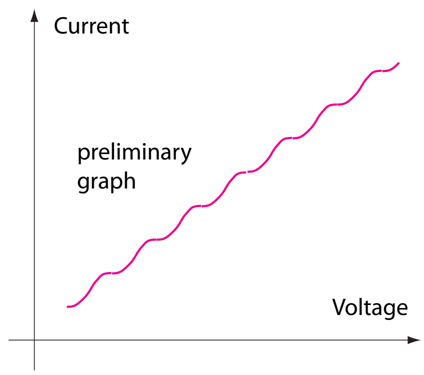


Figure 292 Steps in the flow of electricity in metals

Many other experiments confirm the result and leave only one conclusion: there is a *smallest* charge in nature. This smallest charge has the same value as the charge of an electron. Indeed, electrons turn out to be part of every atom, in a complex way to be explained shortly. In metals, quite a a number of electrons can move freely; that is the reason that metal conduct electricity so well.

Also the flow of matter shows smallest units. We mentioned in the first part that a consequence of the particle structure of liquids is that even in the smoothest of pipes, even oil or any other smooth liquid still produces noise when it flows through the pipe. We mentioned that the noise we hear in our ears in situations of absolute silence, such as in a snowy landscape in the mountains, is due to the granularity of matter. Depending on the material, the smallest units of matter are called atoms, ions or molecules.

Quantons

Electrons, ions, atoms and molecules are *quantum* particles or *quantons*. Like photons, they show some of the aspects of everyday particles, but show many other aspects which are different from what is expected from little stones. Let us have a rapid tour.

Everyday matter has mass, position and momentum, orientation and angular momentum, size, shape, structure and colour. What about matter quantons? First of all, matter quantons do have mass. Single particles can be slowed down or accelerated; in addition, hits by single electrons, atoms or molecules can be detected. Experiments also show that (composed) quantons have structure, size, shape and colour. We will discuss their details below. How do they move while respecting the quantum of action?

Ref. 654 * 'Motion is matter's way of being.'

Figure to be included

Figure 293 Matter diffracts and interferes

Figure to be included

Figure 294 Trying to measure position and momentum

The motion of quantons – matter as waves

Ref. 678 In 1923 and 1924, the French physicist Louis de Broglie pondered over the concept of photon and the possible consequences of the quantum of action for matter particles. It dawned to him that like light quanta, streams of matter particles with the same momentum should also behave as waves. The quantum of action implies wave behaviour. This, de Broglie reasoned, should also apply to matter. He predicted that constant matter flows should have a wavelength and angular frequency given by

$$\lambda = \frac{2\pi\hbar}{p} \quad \text{and} \quad \omega = \frac{E}{\hbar} . \tag{487}$$

where p and E are the momentum and the energy of the single particles. Soon after the prediction, experiments started to provide the confirmation of the idea. It is indeed found that matter streams can diffract, refract and interfere. Due to the small value of the wavelength, one needs careful experiments to detect the effects. Nevertheless, one after the other, *all* experiments which proved the wave properties of light have been repeated for matter beams. For example, in the same way that light diffracts when passing around an edge or through a slit, matter has been found to diffract in these situations. Similarly, researchers inspired by light interferometers built *matter* interferometers; they Ref. 679 work with beams of electrons, nucleons, nuclei, atoms and even large molecules. In the Page 518 same way that the observations of interference of light proves the wave property of light, the interference patterns observed with these instruments show the wave properties of matter.

Like light, matter is also made of particles; like light, matter behaves as a wave when large numbers of particles with the same momentum are involved. Even though beams of large molecules behave as waves, for everyday objects, such as cars on a highway, one never makes such observations. There are two main reasons. First, for cars on highways the involved wavelength is extremely small. Second, the speeds of cars vary too much; streams of objects with the same speed for all objects – only such streams have a chance to be coherent – are extremely rare in nature.

If matter behaves like a wave, we can draw a strange conclusion. For every type of wave, the position X of its maximum and the wavelength λ cannot both be sharply defined simultaneously; on the contrary, their indeterminacies follow the relation

$$\Delta\lambda\Delta X = \frac{1}{2} . \tag{488}$$

Similarly, for every wave the frequency ω and the instant T of its peak amplitude cannot both be sharply defined. Their indeterminacies are related by

$$\Delta\omega\Delta T = \frac{1}{2} . \tag{489}$$

Using the wave properties of matter we get

$$\Delta p\Delta X \geqslant \frac{\hbar}{2} \qquad \text{and} \qquad \Delta E\Delta T \geqslant \frac{\hbar}{2} . \tag{490}$$

These famous relations are called *Heisenberg's indeterminacy relations*. They were discovered by the German physicist Werner Heisenberg in 1925. They state that there is no way to ascribe a precise momentum and position to a material system, nor a precise energy and age. The more accurately one quantity is known, the less accurately the other is.* Matter quantons – like stones, but in contrast to photons – *can* be localized, but only approximately.

Both indeterminacy relations have been checked experimentally in great detail. The limits are easily experienced in experiments. Some attempts are shown in Figure 294. In fact, every experiment proving that matter behaves like a wave is a confirmation of the indeterminacy relation, and vice versa.

As a note, Niels Bohr called the relation between two variables linked in this way *complementarity*. He then explored systematically all such possible pairs. You should search for such observable pairs yourself. Bohr was deeply convinced of the existence of a complementarity principle. Bohr extended this also to philosophical aspects. In a famous story, somebody asked him what was the quantity complementary to precision. He answered: 'Clarity'.

Challenge 1169 ny

In summary, we conclude that the quantum of action prevents position and momentum values to be exactly defined for microscopic systems. Their values are fuzzy. Like Bohr, we will explore some additional limits on motion that follow from the quantum of action.

* The policeman stops the car being driven by Werner Heisenberg. 'Do you know how fast you were driving?' 'No, but I know exactly where I am!'

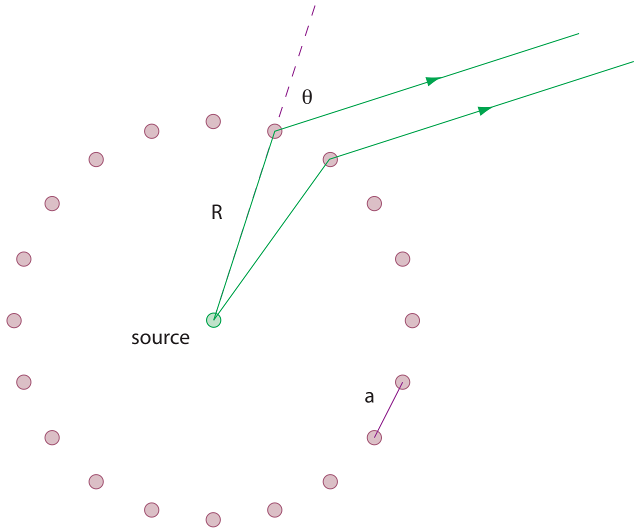


Figure 295 On the quantization of angular momentum

Rotation and the lack of North Poles

Tristo quel discepolo che non avanza il suo maestro.
Leonardo da Vinci*

The quantum of action also has important consequences for rotational motion. Action and angular momentum have the same physical dimensions. It only takes a little thought to show that if matter or radiation has a momentum and wavelength related by the quantum of action, then angular momentum is fixed in multiples of the quantum of action; angular momentum is thus quantized.

Ref. 688 The argument is due to Dicke and Wittke. Just imagine a source at the centre of a circular fence, made of N steel bars spaced by a distance $a = 2\pi R/N$, as shown in Figure 295. In the centre of the fence we imagine a source of matter or radiation that emits particles towards the fence in any chosen direction. The linear momentum of the particle is $p = \hbar k = 2\pi\hbar/\lambda$. Outside the fence, the direction of the particle is given by the condition of positive interference. In other words, the angle θ is given by $a \sin\theta = M\lambda$, where M is an integer. In this process, the fence receives a linear momentum $p \sin\theta$, or an angular momentum $L = pR \sin\theta$. Inserting all expressions one finds that the transferred angular momentum is

$$L = NM\hbar \; . \qquad (491)$$

In other words, the angular momentum of the fence is an integer multiple of $\hbar$. Of course, this argument is only a hint, not a proof. Nevertheless, the argument is correct. The angular momentum of bodies is always a multiple of $\hbar$. Quantum theory thus states that every object rotates in steps.

* 'Sad is that disciple who does not surpass his master.' The statement is painted in large letters in the Aula Magna of the University of Rome.

But rotation has more interesting aspects. Due to the quantum of action, in the same way that linear momentum is fuzzy, angular momentum is so as well. There is an indeterminacy relation for angular momentum L. The complementary variable is the phase angle φ of the rotation. The indeterminacy relation can be expressed in several ways. The simplest, but also the less precise is

Ref. 680
Ref. 681
Page 673

$$\Delta L \, \Delta\varphi \geqslant \frac{\hbar}{2} \,. \tag{492}$$

(The approximation is evident: the relation is only valid for large angular momenta; the relation cannot be valid for small values, as $\Delta\varphi$ by definition cannot grow beyond 2π. In particular, angular momentum eigenstates have $\Delta L = 0$.*)

The quantization and indeterminacy of angular momentum has important consequences. Classically speaking, the poles of the Earth are spots which do not move when observed by a non-rotating observer. Therefore at those spots matter would have a defined position and a defined momentum. However, the quantum of action forbids this. There cannot be a North Pole on Earth. More precisely, the idea of a rotation axis is an approximation not valid in general.

Even more interesting are the effects of the quantum of action on microscopic particles, such as atoms, molecules or nuclei. To begin with, we note that action and angular momentum have the same units. The precision with which angular momentum can be measured depends on the precision of the rotation angle. But if a *microscopic* particle rotates by an angle, this rotation might be unobservable, a situation in fundamental contrast with the case of *macroscopic* objects. Experiments indeed confirm that many microscopic particles have unobservable rotation angles. For example, in many, but not all cases, an atomic nucleus rotated by half a turn cannot be distinguished from the unrotated nucleus.

If a microscopic particle has a *smallest* unobservable rotation angle, the quantum of action implies that the angular momentum of that particle *cannot* be zero. It must always be rotating. Therefore we need to check for each particle what its smallest unobservable angle of rotation is. Physicists have checked experimentally all particles in nature and have found – depending on the particle type – the following smallest unobservable angle values: 0, 4π, 2π, $4\pi/3$, π, $4\pi/5$, $2\pi/3$, etc.

Let us take an example. Certain nuclei have a smallest unobservable rotation angle of *half* a turn. That is the case for a nucleus that looks like a rugby ball and turns around the short axis. Both the largest observable rotation and the indeterminacy are thus a *quarter* turn. Since the change or action produced by a rotation is the number of turns times the angular momentum, we find that the angular momentum of this nucleus is $2 \cdot \hbar$.

As a general result we deduce from the smallest angle values that the angular momentum of a microscopic particle can be 0, $\hbar/2$, $\hbar$, $3\hbar/2$, $2\hbar$, $5\hbar/2$, $3\hbar$, etc. In other words, the intrinsic angular momentum of particles, usually called their *spin*, is an integer mul-

* An exact way to state the indeterminacy relation for angular momentum is

$$\Delta L \, \Delta\varphi \geqslant \frac{\hbar}{2} |1 - 2\pi P(\pi)| \,, \tag{493}$$

where $P(\pi)$ is the normalized probability that the angular position has the value π. For an angular momentum eigenstate, one has $\Delta\varphi = \pi/\sqrt{3}$ and $P(\pi) = 1/2\pi$. This expression has been tested experimentally.

Ref. 682

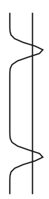


Figure to be included

Figure 296 The Stern–Gerlach experiment

tiple of $\hbar/2$. Spin describes how a particle behaves under rotations. (It turns out that all spin 0 particles are composed and contain other particles; the quantum of action thus remains the limit for rotational motion in nature.)

How can a particle rotate? At this point we do not know yet how to *picture* the rotation. But we can *feel* it. This is done in the same way we showed that light is made of rotating entities: all matter, including electrons, can be *polarized*. This was shown most clearly by the famous Stern–Gerlach experiment.

Silver, Stern and Gerlach

In 1922, Otto Stern and Walter Gerlach* found that a beam of silver atoms that is extracted from an oven splits into two *separate* beams when it passes through an inhomogeneous Ref. 675 magnetic field. There are no atoms that leave the magnetic field in intermediate locations.

The split into two is an intrinsic property of silver atoms. In fact, it is due to their spin value. Silver atoms have spin $\hbar/2$, and depending on its orientation in space, they are deflected either upwards or downwards. There are no intermediate values. The split is a pure quantum effect. This result is so peculiar that it was studied in great detail.

When one of the two beams is selected – say the 'up' beam – and passed through a second set-up, all atoms end up in the 'up' beam. The other exit, for the 'down' beam, remains unused in this case. The up and down beams, in contrast to the original beam, cannot be split further.

But if the second set-up is rotated by $\pi/2$ with respect to the first, again two beams – 'right' and 'left' – are formed; it plays no role whether the incoming beam was from the oven or an 'up' beam. A partially rotated set-up yields a partial, uneven split. The number ratio depends on the angle.

We note directly that if a beam from the oven is split first vertically and then horizontally, the result differs from the opposite order. Splitting processes do not commute. (Whenever the order of two operations is important, physicists speak of lack of 'commutation'.) Since all measurements are processes as well, we deduce that measurements in quantum systems do not commute in general.

Beam splitting is direction dependent. Matter beams behave almost in the same way as polarized light beams. The inhomogeneous magnetic field acts somewhat like a polarizer. The up and down beams, taken together, behave like a fully polarized light beam. In fact,

* Otto Stern (1888–1969) and Walter Gerlach (1889–1979), both German physicists, worked together at the University in Frankfurt.

the polarization direction can be rotated (with the help of a *homogeneous* magnetic field). Indeed, a rotated beam behaves in a unrotated magnet like an unrotated beam in a rotated magnet.

The 'digital' split forces us to rethink the description of motion. In special relativity, the existence of a maximum speed forced us to introduce the concept of space-time, and then to refine the description of motion. In general relativity, the maximum force obliged us to introduce the concepts of horizon and curvature, and then to refine the description of motion. At this point, the existence of the quantum of action forces us to take two similar steps. We will introduce the new concept of Hilbert space, and then we will refine the description of motion.

The language of quantum theory and its description of motion

In classical physics, a physical system is said to *have* momentum and position. The quantum of action makes it impossible to continue using this expression. In classical physics, the state and the measurement result do not need to be distinct, because measurements can be imagined to disturb the system as little as possible. But due to a smallest action in nature, the interaction necessary to perform the measurement cannot be made arbitrarily small. For example, the Stern–Gerlach experiments shows that the measured spin orientation values – like those of any other observable – are *not* intrinsic values, but result from the measurement process itself. The quantum of action thus obliges us to distinguish three entities:

- the state of the system;
- the operation of measurement;
- the result of the measurement.

A general state of a quantum system is thus *not* described by the outcomes of a measurement. The simplest case showing this is the system made of a single particle in the Stern–Gerlach experiment. The experiments showed that a spin measurement on a general (oven) particle state sometimes gives 'up', sometimes 'down' ('up' might be +1, 'down' might be −1) showing that a general state has no properties. It was also found that feeding 'up' into the measurement apparatus gives 'up' states; thus certain special states ('eigenstates') do remain unaffected. Finally, states can be rotated by applied fields; they have an abstract direction.

These details can be formulated in a straightforward way. Since measurements are operations that take a state as input and produce as output a measurement result and an output state, we can say:

- states are described by complex *vectors* in some abstract space called a *Hilbert space*;
- measurements are described by (self-adjoint or) *Hermitean operators* (or matrices);
- measurement results are *numbers*;
- changes of viewpoint are described by *unitary operators* (or matrices) that transform states and measurement operators.

As required by experiment, we have thus distinguished the quantities that are not distinguished in classical physics. Once this step is completed, quantum theory follows quite simply, as we shall see.

Quantum theory describes observables as operators, thus as *transformations* in Hilbert space, because any measurement is an interaction with a system and thus a transformation of its state.

Quantum mechanical experiments also show that the measurement of an observable can only give as result one of the possible eigenvalues of this transformation. The resulting states are eigenvectors – the 'special' states just mentioned. Therefore every expert on motion must know what an eigenvalue and an eigenvector is. For any linear transformation T, those special vectors ψ that are transformed into multiples of themselves,

$$T\psi = \lambda\psi \tag{494}$$

are called *eigenvectors*, and the multiplication factor λ is called the associated *eigenvalue*. Experiments show that the state of the system *after* the measurement is given by the eigenvector of the measured eigenvalue.

The quantum of action obliges us to distinguish between three concepts that are all mixed up in classical physics: the state of a system, the measurement on a system and the measurement result. The quantum of action thus forces us to change the vocabulary with which we describe nature and obliges to use more differentiated concepts. The main step now follows; we describe motion with these concepts. This description is called quantum theory.

In classical physics, motion is given by the path that minimizes the action. Motion takes place in such a way that the action variation δS vanishes when paths with fixed end points are compared. In quantum theory, we need to translate the concept of action and to find a description of variation that is not based on paths, as the concept of 'path' does not exist.

The action variation δS between an initial and a final state is easily defined as

$$\delta S = \langle\psi_\mathrm{i}|\,\delta \int L\mathrm{d}t|\psi_\mathrm{f}\rangle\ , \tag{495}$$

where L is the *Lagrangian (operator)*. The variation of the action is defined in the same way as in classical physics, except that the momentum and position variables are replaced by the corresponding operators.*

In the classical principle of least action, the path is varied while keeping the end points fixed. This variation must be translated into the language of quantum theory. In quantum theory, paths do not exist, because position is not a well-defined observable. The only variation that we can use is

$$\delta\langle\psi_\mathrm{i}|\psi_\mathrm{f}\rangle\ . \tag{496}$$

This complex number describes the variation of the temporal evolution of the system.

The variation of the action must be as small as possible when the temporal evolution is varied, while at the same time it must be impossible to observe actions below $\hbar/2$. This double condition is realized by the so-called *quantum action principle*

$$\langle\psi_\mathrm{i}|\delta \int L\mathrm{d}t|\psi_\mathrm{f}\rangle = -i\hbar\delta\langle\psi_\mathrm{i}|\psi_\mathrm{f}\rangle\ . \tag{497}$$

* More precisely, there is also a condition for ordering of operators in mixed products, so that the lack of commutation between operators is taken into account. We do not explore this issue here.

This principle describes all quantum motion in nature. Classically, the right hand side is zero – since $\hbar$ can then be neglected – and we then recover the minimum action principle of classical physics. In quantum theory however, the variation of the action is proportional to the variation at the end points. The intermediate situations – the 'paths' – do not appear. In the quantum action principle, the factor $-i$ plays an important role. We recall that states are vectors. The factor $-i$ implies that in the complex plane, the complex variation on the right hand side is rotated by a right angle; in this way, even if the variation at the end points is small, no action change below $\hbar/2$ can be observed.

To be convinced about the correctness of the quantum action principle, we proceed in the following way. We first deduce evolution equations, we then deduce all experimental effects given so far, and finally we deduce new effects that we compare to experiments.

The evolution of operators

Since quantum theory distinguishes between states and measurement operators, there are several ways to state equations of motion in quantum theory: above all, we can focus either on the operators or on the states. If we focus on the operators, we get that for a general operator F

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} - \frac{i}{\hbar}[F,H] \; . \qquad (498)$$

This *Heisenberg's equation* of motion for quantum operators. In particular, the Hamiltonian itself follows

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} \; . \qquad (499)$$

If the Hamiltonian does not change under time displacements, like in the case of atoms or any other closed system, the right hand side is zero. This equation thus expresses the conservation of energy.

– CS – more to be written – CS –

The state – or wave function – and its evolution

We can also focus on the change of *states* with time. The quantum action principle then gives

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = H|\psi\rangle \; . \qquad (500)$$

Ref. 683 This famous equation is *Schrödinger's equation of motion*.* In fact, Erwin Schrödinger had

* Erwin Schrödinger (b. 1887 Vienna, d. 1961 Vienna) was famous for being a physicien bohémien, and always lived in a household with two women. In 1925 he discovered the equation which brought him international fame and the Nobel prize for physics in 1933. He was also the first to show that the radiation discovered by Victor Hess in Vienna was indeed coming from the cosmos. He left Germany and then again Austria out of dislike of national socialism, and was for many years professor in Dublin. There he published the famous and influential book *What is life?*. In it, he comes close to predicting the then unknown nuclear acid DNA from theoretical insight alone.

found his equation in two slightly different ways. In his first paper, he used a variational Ref. 684 principle slightly different from the one just given. In the second paper he simply asked: how does the state evolve? He imagined the state of a quanton to behave like a wave and like a particle at the same time. If the state behaves ψ like a wave, it must be described by a function (hence he called it 'wave function') with amplitude W multiplied by a phase factor $e^{i\mathbf{kx}-\omega t}$. The state can thus be written as

$$|\psi\rangle = \psi(t,x) = W(t,x)e^{i\mathbf{kx}-\omega t} . \tag{501}$$

At the same time, the state must also behave like a a particle. In particular, the non-relativistic particle relation between energy and momentum $E = \mathbf{p}^2/2m + V(x)$ must remain valid for these waves. Using the two relations for matter wavelength and frequency, we thus must have

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{\Delta\psi}{2m} + V(x)\psi = H\psi . \tag{502}$$

This 'wave' equation for the complex field ψ became instantly famous in 1926 when Schrödinger, by inserting the potential felt by an electron near a proton, explained the energy levels of the hydrogen atom. In other words, the equation explained the discrete Page 704 colours of all radiation emitted by hydrogen. We will do this below. The frequency of the light emitted by hydrogen gas was found to be in agreement with the prediction of the equation to five decimal places. The aim of describing motion of matter had arrived at a new high point.

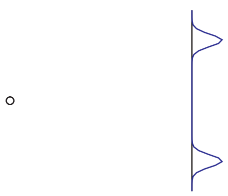
The most precise description of matter is found when the *relativistic* energy–momentum relation is taken into account. We explore this approach below. Even today, predictions of atomic spectra are the most precise and accurate in the whole study of nature. No other description of nature has achieved a higher accuracy.

We delve a bit into the details of the description with the Schrödinger equation (502). The equation expresses a simple connection: the classical speed of matter is the group velocity of the field ψ. We know from classical physics that the group velocity is not always well defined; in cases where the group dissolves in several peaks the concept of group velocity is not of much use; these are the cases in which quantum motion is much different from classical motion, as we will soon discover.

As an example, the left corner pictures in these pages show the evolution of a wave-function – actually its modulus $|\Psi|^2$ – for the case of an exploding two-particle system.

The Schrödinger equation makes another point: velocity and position of matter are not independent variables and cannot be chosen at leisure. Indeed, the initial condition of a system is given by the initial value of the wavefunction alone. No derivatives have to or can be specified. In other words, quantum systems are described by a *first order* evolution equation, in strong contrast to classical physics.

We note for completeness that in the Schrödinger equation the wave function is indeed a vector, despite the apparent differences. The scalar product of two wave functions/vectors is the spatial integral of the product between complex conjugate of the first function and the second function. In this way, all concepts of vectors, such as unit vectors, null vectors, basis vectors, etc. can be reproduced.



Why are atoms not flat? Why do shapes exist?

In 1901, Jean Perrin, and in 1904 the Japanese physicist Nagaoka Hantaro proposed that atoms are small solar systems. In 1913, Niels Bohr used this idea and based his epoch-making atomic calculations on it. All thus somehow assumed that hydrogen atoms are flat. However, this is not observed. Ref. 676 Ref. 677

Atoms, in contrast to solar systems, are *quantum* systems. Atoms, like protons and many other quantum systems, do have sizes and shapes. Atoms are spherical, molecules have more complex shapes. Quantum theory gives a simple recipe for the calculation: the shape of an atom or a molecule is due to the probability distribution of its electrons. The probability distribution is given by the square modulus $|\Psi|^2$ of the wave function.

In other words, Schrödinger's equation defines the shape of molecules. That is why it was said that the equation contains all of chemistry and biology. The precise shape of matter is determined by the interactions of electrons and nuclei. We come back to the issue later. Page 780

In short, the wave aspect of quantons is responsible for all shapes in nature. For example, only the wave aspect of matter, and especially that of electrons, allows to understand the shapes of molecules and therefore indirectly the shapes of all bodies around us, from flowers to people.

Obviously, the quantum of action also implies that shapes fluctuate. If a long molecule is held fixed at its two ends, the molecule cannot remain at rest in between. Such experiments are common today; they confirm that rest does not exist, as it would contradict the existence of a minimum action in nature.

Rest: spread and the quantum Zeno effect

In special relativity, anything moving inertially is at rest. However, the quantum of action implies that no particle can ever be at rest. Therefore, no quantum system can be at in inertial motion. That is the reason that any wave function spreads out in time. In this way, a particle is never at rest, whatever the observer may be.

Only if a particle is bound, not freely moving, one can have the situation that the density distribution is stationary in time.

Another apparent case of rest in quantum theory is called the quantum Zeno effect. Usually, observation changes the system. However, for certain systems, observation can have the opposite effect.

The quantum Zeno effect was partially observed by Wayno Itano and his group in 1990, and definitively observed by Mark Raizen and his group, in 2001. Ref. 686

– CS – more to be told – CS –

Ref. 687 In an fascinating prediction, Saverio Pascazio and his team have predicted that the quantum Zeno effect can be used to realize X-ray tomography of objects with the lowest radiation levels imaginable.

Tunnelling, hills and limits on memory

A slow ball cannot roll over a high hill, says everyday experience. More precisely, classical physics says that if the kinetic energy T is smaller than the potential energy V the ball would have at the top of the hill, the ball cannot roll over the hill. Quantum theory simply states the opposite. There is a probability to pass the hill for *any* energy of the ball.

Since hills in quantum theory are described by potential barriers, and objects by wavefunctions, the effect that an object can pass the hill is called the *tunnelling effect*. For a potential barrier of finite height, any initial wavefunction will spread beyond the barrier. The wavefunction will even be non-vanishing at the place of the barrier. All this is different from everyday experience and thus from classical mechanics.

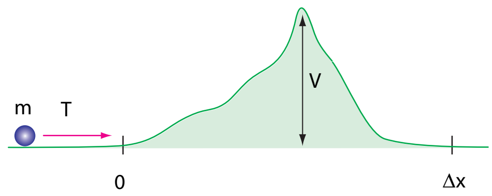


Figure 297 Climbing a hill

Something new is contained in this description of hills: the assumption that all obstacles in nature can be overcome with a *finite* effort. No obstacle is infinitely difficult to surmount. (Only for a potential of infinite height the wavefunction would vanish and not spread on the other side.)

Challenge 1170 ny

How large is the effect? A simple calculation shows that the transmission probability P is given by

$$P \approx \frac{16T(V-T)}{V^2} e^{-2w\sqrt{2m(V-T)/\hbar^2}} \qquad (503)$$

where w is the width of the hill. For a system of many particles, the probability is the product of the probabilities for each particle. In the case of a car in a garage, assuming it is made of 10^{28} atoms of room temperature, and assuming that a garage wall has a thickness of 0.1 m and a potential height of 1 keV= 160 aJ for the passage of an atom, one gets a probability of finding the car outside the garage of

$$P \approx \left(10^{-(10^{12})}\right)^{(10^{28})} \approx 10^{-(10^{40})} . \qquad (504)$$

Challenge 1171 ny

Challenge 1172 ny

This rather small value – just try to write it down to be convinced – is the reason why it is never taken into account by the police when a car is missing. (Actually, the probability is considerably smaller; can you name at least one effect that has been forgotten in this simple calculation?) Obviously, tunnelling can be important only for small systems, made of a few particles, and for thin barriers, with a thickness of the order of $\sqrt{\hbar^2/2m(V-T)}$. Tunnelling of single atoms is observed in solids at high temperature, but is not of importance in daily life. For electrons the effect is larger; the formula gives

$$w \approx 0.5\,\text{nm}\,\sqrt{aJ}\sqrt{V-T} . \qquad (505)$$

At room temperature, kinetic energies are of the order of 6 zJ; increasing temperature obviously increases tunnelling. As a result, electrons or other light particles tunnel quite easily. Indeed, every tv tube uses tunnelling at high temperature to generate the electron beam producing the picture. The heating is the reason that tv tubes take time to switch on.

For example, the tunnelling of electrons limits the ability to reduce the size of computer memories, and thus makes it impossible to produce silicon integrated circuits with one Challenge 1173 ny terabyte (TB) of random access memory (RAM). Are you able to imagine why? In fact, tunnelling limits the working of any type of memory, also that of our brain. If we would be much hotter than 37°C, we could not remember anything!

By the way, light, being made of particles, can also tunnel through potential barriers. The best potential barriers for light are called *mirrors*; they have barrier heights of the order of one aJ. Tunnelling implies that light can be detected behind a mirror. These so-called evanescent waves have indeed been detected. They are used in several high-precision experiments.

Spin and motion

Everything turns.

Anonymous

Spin describes how a particle behaves under rotations. The full details of spin of electrons were deduced from experiments by two Dutch students, George Uhleneck and Samuel Ref. 691 Goudsmit. They had the guts to publish what also Ralph Kronig had suspected: that electrons rotate around an axis with an angular momentum of $\hbar/2$. In fact, this value is correct for all elementary matter particles. (In contrast, radiation particles have spin values that are integer multiples of $\hbar$.) In particular, Uhlenbeck and Goudsmit proposed a g-value of 2 for the electron in order to explain the optical spectra. The factor was explained by Ref. 692 Llewellyn Thomas as a relativistic effect a few months afterwards.

In 2004, experimental techniques became so sensitive that the magnetic effect of a single electron spin attached to an impurity (in an otherwise unmagnetic material) has been detected. Researchers now hope to improve these so-called magnetic resonance force microscopes until they reach atomic resolution.

In 1927, the Austrian physicist Wolfgang Pauli* discovered how to include spin in a quantum mechanical description; instead of a state function with a single component, one needs a state function with *two* components. Nowadays, Pauli's equation is mainly of conceptual interest, because like the one by Schrödinger, it does not comply with special relativity. However, the idea to double the necessary components was taken up by Dirac

* Wolfgang Ernst Pauli (b. 1900 Vienna, d. 1958 Zürich), when 21 years old, wrote one of the best texts on special and general relativity. He was the first to calculate the energy levels of hydrogen with quantum theory, discovered the exclusion principle, included spin into quantum theory, elucidated the relation between spin and statistics, proved the CPT theorem and predicted the neutrino. He was admired for his intelligence and feared for his biting criticisms, which lead to his nickname 'conscience of physics'. Despite this habit he helped many people in their research, such as Heisenberg with quantum theory, without claiming any Ref. 693 credit for himself. He was seen by many, including Einstein, as the greatest and sharpest mind of twentieth century physics. He was also famous for the 'Pauli effect', i.e. his ability to trigger disasters in laboratories, machines and his surroundings by his mere presence. As we will see shortly, one can argue that Pauli got the Nobel Prize in physics in 1945 (officially 'for the discovery of the exclusion principle') for finally settling the question on the number of angels that can dance on the tip of a pin.

when he introduced the relativistic description of the electron, and the idea is used for all other particle equations.

Relativistic wave equations

A few years after Max Planck had discovered the quantum of action, Albert Einstein published the theory of special relativity. The first question Planck asked himself was whether the quantum of action would be independent of the observer. For this reason, he invited Einstein to Berlin. Doing so, he made the then unknown patent office clerk famous in physicist circles.

The quantum of action is indeed independent of the speed of the observer. All observers find the same minimum value. As a result, the evolution equation (498) for operators is valid also when the constancy of the speed of light is taken into account. To include special relativity into quantum theory, we only need to find the correct quantum Hamiltonian operator.

Given that the classical Hamiltonian of a free particle is given by

$$H = \beta\sqrt{c^4 m^2 + c^2 \mathbf{p}^2} \quad \text{with} \quad \mathbf{p} = \gamma m \mathbf{v} \,, \tag{506}$$

one might ask: what is the corresponding Hamilton operator? A simple answer was given, only in 1950, by L.L. Foldy and S.A. Wouthuysen. The operator is almost the same one: Ref. 699

$$H = \beta\sqrt{c^4 m^2 + c^2 \mathbf{p}^2} \quad \text{with} \quad \beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{507}$$

The signs of the operator β distinguishes particles and antiparticles; it has two 1s and two -1s to take care of the two possible spin directions. With this Hamiltonian operator, a wavefunction for a particle has vanishing antiparticle components, and vice versa. The Hamilton operator yields the velocity operator $\mathbf{v}$ through the same relation that is valid in classical physics:

$$\mathbf{v} = \frac{d}{\mathrm{d}t}\mathbf{x} = \beta \frac{\mathbf{p}}{\sqrt{c^4 m^2 + c^2 \mathbf{p}^2}} \,. \tag{508}$$

This velocity operator shows a continuum of eigenvalues from minus to plus the speed of light. The velocity $\mathbf{v}$ is a constant of motion, as are $\mathbf{p}$ and $\sqrt{c^4 m^2 + c^2 \mathbf{p}^2}$.

The orbital angular momentum $\mathbf{l}$ is also defined as in classical physics through

$$\mathbf{l} = \mathbf{x} \times \mathbf{p} \,. \tag{509}$$

Ref. 701 Both the orbital angular momentum $\mathbf{l}$ and the spin σ are separate constants of motion. A particle (or antiparticle) with positive (or negative) component has a wave function with only one non-vanishing component; the other three components vanish.

But alas, the representation of relativistic motion given by Foldy and Wouthuysen is

not the most simple for a generalization to particles when electromagnetic interactions are present. The simple identity between classical and quantum-mechanical description is lost when electromagnetism is included. We give below the way to solve the problem.

Maximum acceleration

Combining quantum theory with special relativity leads to a maximum acceleration value for microscopic particles. Using the time–energy indeterminacy relation, you can deduce
Challenge 1174 n
that

$$a \leqslant \frac{2mc^3}{\hbar} . \tag{510}$$

Up to the present, no particle has ever been observed with a higher acceleration than
Ref. 689
this value. In fact, no particle has ever been observed with accelerations approaching this value. We note that the acceleration limit is different from the acceleration limit due to general relativity:

$$a \leqslant \frac{c^4}{4Gm} . \tag{511}$$

In particular, the quantum limit (510) applies to microscopic particles, whereas the general relativistic limit applies to macroscopic systems. Can you confirm that in each domain the respective limit is the smaller of the two?
Challenge 1175 e

– CS – the rest of quantum theory will appear in the next version – CS –

the examples so far have shown how quantons move that are described only by mass. Now we study the motion of quantum systems that are electrically charged.

Curiosities and fun challenges about quantum theory

Quantum theory is so full of strange results that all of it could be titled 'curiosities'. A few of the prettier cases are given here.

- The quantum of action implies that there are no fractals in nature. Can you confirm this result?
Challenge 1176 ny

- Can atoms rotate? Can an atom that falls on the floor roll under the table? Can atoms be put into high speed rotation? The answer is no to all questions, because angular mo-
Ref. 704
mentum is quantized and because atoms are not solid objects. The macroscopic case of an object turning slower and slower until it stops does not exist in the microscopic world.

Challenge 1177 ny

Can you explain how this follows from the quantum of action?

Ref. 707

Challenge 1178 n

▪ Do hydrogen atoms exist? Most types of atoms have been imaged with microscopes, photographed under illumination, levitated one by one, and even moved with needles, one by one, as the picture shows. Others have moved single atoms using laser beams to push them. However, not a single of these experiments measured hydrogen atoms. Is that a reason to doubt the existence of hydrogen atoms? Taking seriously this not-so-serious discussion can be a lot of fun.

Ref. 706

▪ Light is refracted when entering dense matter. Do matter waves behave similarly? Yes, they do. In 1995, David Pritchard showed this for sodium waves entering helium and xenon gas.

Challenge 1179 ny

▪ Two observables can commute for two different reasons: either they are very *similar*, such as the coordinate x and x^2, or they are very *different*, such as the coordinate x and the momentum p_y. Can you give an explanation?

Challenge 1180 ny

▪ Space and time translations commute. Why then do the momentum operator and the Hamiltonian not commute in general?

Ref. 708

Challenge 1181 ny

▪ With two mirrors and a few photons, it is possible to capture an atom and keep it floating between the two mirrors. This feat, one of the several ways to isolate single atoms, is now standard practice in laboratories. Can you imagine how it is realized?

Ref. 709

▪ For a bound system in a non-relativistic state with no angular momentum, one has the relation

$$\langle rr \rangle \langle T \rangle \geqslant \frac{9\hbar^2}{8m} , \qquad (512)$$

Challenge 1182 n

where m is the reduced mass and T the kinetic energy of the components, and r the size of the system. Can you deduce the result and check it for hydrogen?

Challenge 1183 n

▪ Electrons don't like high magnetic fields. When a magnetic field is too high, electrons are squeezed into a small space, in the direction transversal to their motion. If this spacing becomes smaller than the Compton wavelength, something special happens. Electron-positron pairs appear from the vacuum and move in such a way as to reduce the applied magnetic field. The corresponding field value is called the quantum *critical magnetic field*. Physicists also say that the Landau levels spacing then becomes larger than the electron rest energy. Its value is about 4.4 GT. Nevertheless, in magnetars, fields over 20 times as high have been measured. How is this possible?

Challenge 1184 ny

▪ Often one reads that the universe might have been born from a quantum fluctuation. Does this statement make sense?

22. Colours and other interactions between light and matter

After the description of the motion of matter and radiation, the next step is the description of their interactions. In other words, how do charged particles react to electromagnetic fields and vice versa? Interactions lead to surprising effects, most of which appear when the problem is treated taking special relativity into account.

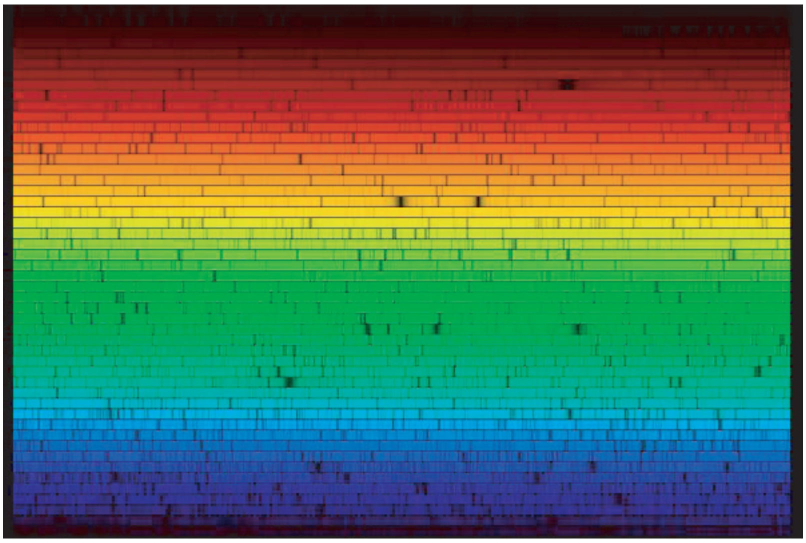


Figure 298 The spectrum of daylight: a stacked section of the rainbow (© Nigel Sharp (noao), fts, nso, kpno, aura, nsf)

What are stars made of?

In the beginning of the eighteenth century the English physicist William Wollaston and again the Bavarian instrument maker Joseph Fraunhofer* noted that the rainbow *lacks* certain colours. These colours appear as black lines when the rainbow is spread out sufficiently. Figure 298 shows them in detail; the lines are called *Fraunhofer lines* today. In 1860, Gustav Kirchhoff and Robert Bunsen showed that the missing colours were exactly those colours that certain elements emitted when heated. With a little of experimenting they managed to show that sodium, calcium, barium, nickel, magnesium, zinc, copper and iron existed on the Sun. However, they were unable to attribute 13 of the 476 lines they observed. In 1868, Jules Janssen and Joseph Lockyer independently predicted that the unknown lines were from a new element; it was eventually found also on Earth, in an uranium mineral called cleveite, in 1895. Obviously it was called 'helium', from the Greek word 'helios' – Sun. Today we know that it is the second ingredient of the Sun, in order of frequency, and of the universe, after hydrogen. Helium, despite being so common, is rare on Earth because it is a light noble gas that does not form chemical components. Helium thus tends to rise in the atmosphere until it leaves the Earth.

* Born as Joseph Fraunhofer (b. 1787 Straubing, d. 1826 München). Bavarian, orphan at 11, he learned lens polishing at that age; autodidact, he studied optics from books. He entered an optical company at age 19, ensuring the success of the business, by producing the best available lenses, telescopes, micrometers, optical gratings and optical systems of his time. He invented the spectroscope and the heliometer. He discovered and counted 476 lines in the spectrum of the Sun, today named after him. Up to this day, Fraunhofer lines are used as measurement standards. Physicists across the world would buy their equipment from him, visit him and ask for copies of his publications. Even after his death, his instruments remain unsurpassed. With his telescopes, in 1837 Bessel was able to measure the first parallax of a star and in 1846 Johann Gottfried Galle discovered Neptune. Fraunhofer became professor in 1819; he died young, from the consequences of the years spent working with lead and glass powder.

Understanding the colour lines produced by each element had started to become of interest already before the discovery of helium; the interest rose even more afterwards, due to the increasing applications of colours in chemistry, physics, technology, crystallography, biology and lasers. It is obvious that classical electrodynamics cannot explain the sharp lines. Only quantum theory can explain colours.

What determines the colour of atoms?

The simplest atom to study is the atom of hydrogen. Hydrogen gas emits light consisting of a number of sharp spectral lines. Already in 1885 century the Swiss school teacher Johann Balmer (1828–1898) had discovered that the wavelengths of visible lines follow from the formula

$$\frac{1}{\lambda_{\mathrm{mn}}} = R\left(\frac{1}{4} - \frac{1}{m^2}\right) . \tag{513}$$

This expression was generalized by Johannes Rydberg (1854–1919) to include the ultraviolet and infrared colours

$$\frac{1}{\lambda_{\mathrm{mn}}} = R\left(\frac{1}{n^2} - \frac{1}{m^2}\right) , \tag{514}$$

where n and $m > n$ are positive integers, and the so-called Rydberg constant R has the value $10.97\,\mu\mathrm{m}^{-1}$. Thus quantum theory has a clearly defined challenge here: to explain the formula and the value of R.

By the way, the transition λ_{21} for hydrogen – the shortest wavelength possible – is called the *Lyman-alpha line*. Its wavelength, 121.6 nm, lies in the ultraviolet. It is easily observed with telescopes, since most of the visible stars consist of excited hydrogen. The Lyman-alpha line is regularly used to determine the speed of distant stars or galaxies, since the Doppler effect changes the wavelength when the speed is large. The record so far is a galaxy found in 2004 with a Lyman-alpha line shifted to 1337 nm. Can you calculate the speed with which it moves away from the Earth?

Ref. 694

Challenge 1185 ny

There are many ways to deduce Balmer's formula from the minimum action. In 1926, Schrödinger solved his equation of motion for the electrostatic potential $V(r) = e^2/4\pi\varepsilon_0 r$ of a point-like proton; this famous calculation however, is long and complex. In order to understand hydrogen colours, it is not necessary to solve an equation of motion; it is sufficient to compare the initial and final state. This can be done most easily by noting that a specific form of the action must be a multiple of $\hbar/2$. This approach was developed by Einstein, Brillouin and Keller and is now named after them. It states that the action S of any quantum system obeys

Ref. 695

$$S = \frac{1}{2\pi}\oint \mathrm{d}q_{\mathrm{i}}\, p_{\mathrm{i}} = \left(n_{\mathrm{i}} + \frac{\mu_{\mathrm{i}}}{4}\right)\hbar \tag{515}$$

for every coordinate q_{i} and its conjugate momentum p_{i}. Here, n_{i} can be 0 or any positive integer and μ_{i} is the so-called *Maslov index* (an even integer) that in the case of atoms has the value 2 for the radial and azimuthal coordinates r and θ, and 0 for the rotation angle φ.

Any rotational motion in a spherical potential $V(r)$ is characterized by a constant

Figure to be included

Figure 299 The energy levels of hydrogen

energy E, and constant angular momenta L and L_z. Therefore the conjugate momenta for the coordinates r, θ and φ are

Challenge 1186 ny

$$p_r = \sqrt{2m(E - V(r)) - \frac{L^2}{r^2}}$$
$$p_\theta = \sqrt{L^2 - \frac{L_z^2}{\sin^2\theta}}$$
$$p_\varphi = L_z \; . \qquad (516)$$

Using these expressions in equation (515) and setting $n = n_r + n_\theta + n_\varphi + 1$ yields* the result

$$E_\mathrm{n} = \frac{1}{n^2}\frac{-me^4}{2(4\pi\varepsilon_0)^2\hbar^2} = \frac{-R}{n^2} \approx \frac{2.19\,\mathrm{aJ}}{n^2} \approx \frac{13.6\,\mathrm{eV}}{n^2} \; . \qquad (519)$$

Using the idea that a hydrogen atom emits a single photon when its electron changes from state E_n to E_m, one gets the formula found by Balmer and Rydberg. (This whole discussion assumes that the electrons in hydrogen atoms are in eigenstates. Can you argue why this is the case?)

Challenge 1188 e

Challenge 1189 ny

The effective radius of the electron orbit in hydrogen is given by

$$r_\mathrm{n} = n^2 \frac{\hbar^2 4\pi\varepsilon_0}{\pi m e^2} = n^2 a_0 \approx n^2\, 53\,\mathrm{pm} \; . \qquad (520)$$

The smallest value 53 pm for $n = 1$ is called the *Bohr radius* and is abbreviated a_0. Quantum theory thus implies that a hydrogen atom excited to the level $n = 500$ is

* The calculation is straightforward. After insertion of $V(r) = e/4\pi\varepsilon_0 r$ into equation (516) one needs to perform the (tricky) integration. Using the general result

Challenge 1187 ny

$$\frac{1}{2\pi}\oint \frac{\mathrm{d}z}{z}\sqrt{Az^2 + 2Bz - C} = -\sqrt{C} + \frac{B}{\sqrt{-A}} \qquad (517)$$

one gets

$$(n_r + \frac{1}{2})\hbar + L = n\hbar = \frac{e^2}{4\pi\varepsilon_0}\sqrt{\frac{m}{-2E}} \; . \qquad (518)$$

This leads to the energy formula (519).

about 12 µm in size, larger than many bacteria! This feat has indeed been achieved, even though such blown-up atoms, usually called Rydberg atoms, are extremely sensitive to perturbations.
Ref. 697

The orbital frequency of electrons in hydrogen is

$$f_n = \frac{1}{n^3} \frac{e^4 m}{4\varepsilon_0^2 h^3} \tag{521}$$

and the electron speed is

$$v_n = \frac{e^2}{2n\varepsilon_0 h} \approx \frac{2.2\,\mathrm{Mm/s}}{n} \approx \frac{0.007\,c}{n} \tag{522}$$

As expected, the further electrons orbit the nucleus, the slower they move. This result *can* be checked by experiment. Exchanging the electron by a muon allows to measure the time dilation of its lifetime. The measurements coincide with the formula. We note that the speeds are slightly relativistic. However, this calculation did not take into account relativistic effects. Indeed, precision measurements show slight differences between the calculated energy levels and the measured ones.
Ref. 696

Relativistic hydrogen

Also in the relativistic case, the EBK action has to be a multiple of $\hbar/2$. From the relativistic expression of energy
Ref. 695

$$E + mc^2 = \sqrt{p^2c^2 + m^2c^4} - \frac{e^2}{4\pi\varepsilon_0 r} \tag{523}$$

Challenge 1190 ny

we get the expression

$$p_r^2 = 2mE\left(1 + \frac{E}{2mc^2}\right) + \frac{2me^2}{4\pi\varepsilon_0 r}\left(1 + \frac{E}{mc^2}\right) . \tag{524}$$

We now use the expression for the dimensionless fine structure constant $\alpha = \frac{e^2}{4\pi\varepsilon_0 \hbar c} = \sqrt{2R/mc^2} \approx 1/137.036$. The radial EBK action then implies that
Challenge 1191 ny

$$(E_{nl} + mc^2) = mc^2 \sqrt{1 + \frac{\alpha^2}{\left(n - l - \frac{1}{2} + \sqrt{(l + \frac{1}{2})^2 - \alpha^2}\right)^2}} . \tag{525}$$

This result is correct for point-like electrons. In reality, the electron has spin 1/2; the correct relativistic energy levels thus appear when we set $l = j \pm 1/2$ in the above formula. The result can be approximated by

$$E_{nj} = \frac{-R}{n^2}\left(1 + \frac{\alpha^2}{n^2}\left(\frac{n}{j + \frac{1}{2}} - \frac{3}{4}\right) + ...\right) \tag{526}$$

It reproduces the hydrogen spectrum to an extremely high accuracy. Only the introduction of virtual particle effects yields an even better result. We will present this point later on.

Relativistic wave equations – again

> The equation was more intelligent than I was.
> Paul Dirac about his equation, repeating a statement made by Heinrich Hertz.

Unfortunately, the representation of relativistic motion given by Foldy and Wouthuysen is not the most simple for a generalization to particles in the case that electromagnetic interactions are present. The simple identity between classical and quantum-mechanical description is lost if electromagnetism is included.

Charged particles are best described by another, equivalent representation of the Hamiltonian, which was discovered much earlier, in 1926, by the British physicist Paul Dirac.* Dirac found a neat trick to take the square root appearing in the relativistic energy operator. In this representation, the Hamilton operator is given by

$$H_{\mathrm{Dirac}} = \beta m + \alpha \cdot \mathbf{p} \tag{527}$$

Its position operator x is not the position of a particle, but has additional terms; its velocity operator has only the eigenvalues plus or minus the velocity of light; the velocity operator is not simply related to the momentum operator; the equation of motion contains the famous 'Zitterbewegung' term; orbital angular momentum and spin are not separate constants of motion.

So why use this horrible Hamiltonian? It is the only Hamiltonian that can be easily used for charged particles. Indeed, it is transformed to the one coupled to the electromagnetic field by the so-called minimal coupling, i.e. by the substitution Page 509

$$\mathbf{p} \rightarrow \mathbf{p} - q\mathbf{A} . \tag{528}$$

that treats electromagnetic momentum like particle momentum. With this prescription, Dirac's Hamiltonian describes the motion of charged particles interacting with an electromagnetic field **A**. This substitution is not possible in the Foldy–Wouthuysen Hamiltonian. In the Dirac representation, particles are pure, point-like, structureless electric charges; in the Foldy–Wouthuysen representation they acquire a charge radius and a magnetic moment interaction. (We come back to the reasons below, in the section on QED.) Ref. 700

* Paul Adrien Maurice Dirac (b. 1902 Bristol, d. 1984 Tallahassee), British physicist, born as son of a French-speaking Swiss immigrant. He studied electrotechnics in Bristol, then went to Cambridge, where he later became professor on the chair Newton had held before. In the years from 1925 to 1933 he published a stream of papers, of which several were worth a Nobel prize, which he received in 1933. He unified special relativity and quantum theory, he predicted antimatter, he worked on spin and statistics, he predicted magnetic monopoles, he speculated on the law of large numbers etc. His introversion, friendliness and shyness, his deep insights into nature, combined with a dedication to beauty in theoretical physics, made him a legend all over the world already during his lifetime. For the latter half of his life he tried, unsuccessfully, to find an alternative to quantum electrodynamics, of which he was the founder, as he was repelled by the problems of infinities. He died in Florida, where he lived and worked after his retirement from Cambridge.

In more detail, the simplest description of an electron (or any other elementary, stable, electrically charged particle of spin 1/2) is given by the equations

$$\frac{d\rho}{dt} = [H, \rho]$$

$$H_{\text{Dirac}} = \beta mc^2 + \alpha \cdot (\mathbf{p} - q\mathbf{A}(\mathbf{x}, t))c + q\varphi(\mathbf{x}, t) \quad \text{with}$$

$$\alpha_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \alpha_2 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix} \quad \alpha_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

$$\beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

$$H_{\text{Maxwell}} = \quad . \tag{529}$$

The first Hamiltonian describes how charged particles are moved by electromagnetic fields, and the second describes how fields are moved by charged particles. Together, they form what is usually called *quantum electrodynamics* or QED for short.

As far as is known today, the relativistic description of the motion of charged matter and electromagnetic fields given by equation (529) is *perfect:* no differences between theory and experiment have ever been found, despite intensive searches and despite a high reward for anybody who would find one. All known predictions completely correspond with the measurement results. In the most spectacular cases, the correspondence between theory and measurement is more than fourteen digits. But the precision of QED is less interesting than those of its features that are missing in classical electrodynamics. Let's have a quick tour.

– CS – more to come – CS –

Antimatter

Antimatter is now a household term. Interestingly, the concept was formed *before* any experimental evidence for it was known. Indeed, the antimatter companion of the electron was predicted in 1926 by Paul Dirac from his equation. Without knowing this prediction, Carl Anderson discovered it in 1932 and called it *positron*, even though 'positon', without the 'r', would have been the correct name. Anderson was studying cosmic rays and noticed that some 'electrons' were turning the wrong way in the magnetic field he had applied to his apparatus. He checked everything in his machine and finally deduced that he found a particle with the same mass as the electron, but with positive electric charge.

The existence of positrons has many strange implications. Already in 1928, before their discovery, the swedish theorist Oskar Klein had pointed out that Dirac's equation for electrons makes a strange prediction: when an electron hits a sufficiently steep potential wall, the reflection coefficient is larger than unity. Such a wall will reflect *more* than what is thrown at it. In 1935, after the discovery of the positron, Werner Heisenberg and Hans Euler explained the paradox. They found that the Dirac equation predicts a surprising effect: if an electric field exceeds the critical value of

Ref. 702

$$E_c = \frac{m_e c^2}{e \lambda_e} = \frac{m_e^2 c^3}{e \hbar} = 1.3\,\mathrm{EV/m}\ , \qquad (530)$$

the vacuum will spontaneously generate electron–positron pairs, which then are separated by the field. As a result, the original field is reduced. This so-called *vacuum polarization* is also the reason for the reflection coefficient greater than unity found by Klein, since steep potentials correspond to high electric fields.

Truly gigantic examples of vacuum polarization, namely around charged black holes, will be described later on.

Page 815

We note that such effects show that the *number* of particles is not a constant in the microscopic domain, in contrast to everyday life. Only the *difference* between particle number and antiparticle number turns out to be conserved. This topic will be expanded in the chapter on the nucleus.

Of course, the generation of electron–positron pairs is not a *creation* out of nothing, but a *transformation* of energy into matter. Such processes are part of every relativistic description of nature. Unfortunately, physicists have the habit to call this transformation 'creation' and thus confuse this issue somewhat. Vacuum polarization is a process transforming, as we will see, *virtual* photons into matter. That is not all: the same can also be done with *real* photons.

Virtual particles and QED diagrams

Contrary to what was said so far, there is a case where actions smaller than the minimal one do play a role. We already encountered an example. In the collision between two electrons, there is an exchange of virtual photons. We know that this process cannot be observed. Indeed, the action value S for this exchange obeys

$$S \leqslant \frac{\hbar}{2}\ . \qquad (531)$$

In short, *virtual* particles are those particles that appear only in interactions. Virtual particles are intrinsically short-lived; they are the opposite of free particles. In a sense, virtual particles are *bound* particles; they are bound in space and time.

– CS – more to come – CS –

In summary, all virtual matter and radiation particle-antiparticles pairs together form what we call the vacuum; in addition, virtual radiation particles form static fields. Virtual particles are needed for a full description of interactions, and in particular, they are responsible for every decay process.

We will describe a few more successes of quantum theory shortly. Before we do that, we settle one important question.

Compositeness

When is an object composite? Quantum theory gives several pragmatic answers. The first one is somewhat strange: an object is composite when its gyromagnetic ratio is different than the one predicted by QED. The *gyromagnetic ratio* γ is defined as the ratio between the magnetic moment **M** and the angular momentum **L**. In other terms,

$$\mathbf{M} = \gamma \mathbf{L} . \tag{532}$$

The gyromagnetic ratio γ is measured in $s^{-1}T^{-1}$=C/kg and determines the energy levels of magnetic spinning particles in magnetic fields; it will reappear later in the context of Page 836 magnetic resonance imaging. All candidates for elementary particles have spin 1/2. The gyromagnetic ratio for spin 1/2 particles of mass m can be written as

$$\gamma = \frac{M}{\hbar/2} = g\frac{e}{2m} . \tag{533}$$

(The expression $e\hbar/2m$ is often called the magneton of the particle; the dimensionless factor g/2 is often called the gyromagnetic ratio as well; this sometimes leads to confusion.) The criterion of elementarity thus can be reduced to a criterion on the value of the dimensionless number g, the so-called g-*factor*. If the g-factor *differs* from the value predicted by QED for point particles, about 2.0, the object is *composite*. For example, a Challenge 1192 ny $^4He^+$ helium ion has a spin 1/2 and a g value of $14.7 \cdot 10^3$. Indeed, the radius of the helium ion is $3 \cdot 10^{-11}$ m, obviously finite and the ion is a composite entity. For the proton, one measures a g-factor of about 5.6. Indeed, experiments yield a finite proton radius of about 0.9 fm.

Also the neutron, which has a magnetic moment despite being neutral, must therefore be composite. Indeed, its radius is approximately that of the proton. Similarly, molecules, mountains, stars and people must be composite. Following this first criterion, the only elementary particles are *leptons* – i.e. electrons, muons, tauons and neutrinos –, *quarks* and *intermediate bosons* – i.e. photons, W-bosons, Z-bosons and gluons. More details on Page 836 these particles will be uncovered in the chapter on the nucleus.

Another simple criterion for compositeness has just been mentioned: *any object with a measurable size is composite*. This criterion produces the same list of elementary particles as the first. Indeed, this criterion is related to the previous one. The simplest models for Ref. 703 composite structures predicts that the g-factor obeys

$$g - 2 = \frac{R}{\lambda_C} \tag{534}$$

where R is the radius and $\lambda_C = h/mc$ the Compton wavelength of the system. The expres-

sion is surprisingly precise for helium 4 ions, helium 3, tritium ions and protons, as you
Challenge 1193 e might want to check.

A third criterion for compositeness is more general: *any object larger than its Compton length is composite*. The background idea is simple. An object is composite if one can detect *internal* motion, i.e. motion of some components. Now the action of any part with mass m_{part} moving inside a composed system of size r follows

$$S_{\text{part}} < 2\pi\, r\, m_{\text{part}}\, c < \pi\, r\, m\, c \tag{535}$$

where m is the mass of the *composite* object. On the other hand, following the principle of quantum theory, this action, to be observable, must be larger than $\hbar/2$. Inserting this condition, we find that for any composite object*

$$r > \frac{\hbar}{2\pi\, m\, c} . \tag{536}$$

The right hand side differs only by a factor $4\pi^2$ from the so-called *Compton (wave)length*

$$\lambda = \frac{h}{m\, c} . \tag{537}$$

of an object. Any object *larger* than its own Compton wavelength is thus composite. Any object *smaller* than the right hand side of expression (536) is thus elementary. Again, only leptons, including neutrinos, quarks and intermediate bosons pass the test. All other objects are composite, as the tables in Appendix C make clear. This third criterion produces
Challenge 1195 ny the same list as the previous ones. Can you explain the reason?

Interestingly, the topic is not over yet. Even stranger statements about compositeness will appear when gravity is taken into account. Just be patient; it is worth it.

Curiosities and fun challenges about colour

Colours are at least as interesting in quantum theory as they are in classical electrodynamics.

▪ If atoms contain orbiting electrons, the rotation of the Earth, via the Coriolis acceleration, should have an effect on their motion. This beautiful prediction is due to Mark
Ref. 696 Silverman; the effect is so small however, that is has not been measured yet.

▪ Light is diffracted by material gratings. Can matter be diffracted by light gratings? Surprisingly, it actually can, as predicted by Dirac and Kapitza in 1937. In 1986, this was
Ref. 705 accomplished with atoms. For free electrons the feat is more difficult; the clearest confirmation came in 2001, when the technology advances for lasers were used to perform a beautiful measurement of the typical diffraction maxima for electrons diffracted by a light grating.

▪ Light is totally reflected when it is directed to a dense material under an angle so large that it cannot enter it any more. Interestingly, in the case that the material is excited, the

Challenge 1194 ny * Can you find the missing factor of 2? And is the assumption valid that the components must always be lighter than the composite?

Ref. 696 totally reflected beam can be *amplified*. This has been shown by several Russian physicists.

▪ Where is the sea bluest? Sea water is blue because it absorbs red and green light. Sea water can also be of bright colour if the sea floor reflects light. Sea water is often also green, because it often contains small particles that scatter or absorb blue light. Most frequently, this is soil or plankton. The sea is thus especially blue if it is deep, clear and cold, so that it is low in plankton content. (Satellites determine plankton content from the 'greenness' of the sea.) There is a place where the sea is deep, cold and quiet for most parts of the year: the Sargasso sea. It is often called the bluest spot of the oceans.

The strength of electromagnetism

The great Wolfgang Pauli used to say that after his death, the first question he would ask the devil would be an explanation of Sommerfeld's fine structure constant. (People used to comment that after the devil will have explained it to him, he would think a little, and then snap 'Wrong!') The name *fine structure constant* was given by Arnold Sommerfeld Page 534 to the dimensionless constant of nature given by

$$\alpha = \frac{e^2}{4\pi\varepsilon_0 \hbar c} \approx \frac{1}{137.035\,999\,76(50)} \approx 0.007\,297\,352\,533(27)\ . \qquad (538)$$

This number first appeared in explanations for the fine structure of certain atomic colour spectra, hence its name. Sommerfeld was the first to understand its general importance. The number is central to quantum electrodynamics for several reasons. First of all, it describes the *strength* of electromagnetism. Since all charges are multiples of the electron charge, a higher value would mean a stronger attraction or repulsion between charged bodies. The value of α thus determines the size of atoms, and thus the size of all things, as well as all colours.

Secondly, only because this number is quite a bit smaller than unity are we able to talk about particles at all. The argument is somewhat involved; it will be detailed later on. In any case, only the small value of the fine structure constant makes it possible to distinguish particles from each other. If the number were near or larger than one, particles would interact so strongly that it would *not* be possible to observe or to talk about particles at all.

This leads to the third reason for the importance of the fine structure constant. Since it is a dimensionless number, it implies some yet unknown mechanism that fixes its value. Uncovering this mechanism is one of the challenges remaining in our adventure. As long as the mechanism remains unknown, we do not understand the colour and size of a single thing around us.

Small changes in the strength of electromagnetic attraction between electrons and protons would have numerous important consequences. Can you describe what would happen to the size of people, to the colour of objects, to the colour of the Sun or to the workings of computers if the strength would double? And if it would drop to half the usual value over time? Challenge 1196 ny

Explaining the number is the most famous and the toughest challenge of modern physics since the issue appeared in the 1920s. It is the reason for Pauli's request to the devil. In 1946, during his Nobel Prize lecture, he repeated the statement that a theory that does not

Ref. 710 determine this number cannot be complete. The challenge is so tough that for the first 50 years there were only two classes of physicists: those who did not even dare to take on the challenge, and those who had no clue. This fascinating story still awaits us.

The topic of the fine structure constant is so deep that it leads many astray. For example, it is often heard that in physics it is impossible to change physical units in such a way that $\hbar$, c and e are equal to 1 at the same time; these voices suggest that doing so would not allow that the number 1/137.036... would keep its value. Can you show that the argument is wrong, and that doing so does not affect the fine structure constant? Challenge 1197 n

To continue with the highest efficiency on our path across quantum theory, we first look at two important topics: the issue of indistinguishability and the issue of interpretation of its probabilities.

Bibliography

647 GIUSEPPE FUMAGALLI, *Chi l'ha detto?*, Hoepli, Milano, 1983. Cited on page 656.

648 The quantum of action is introduced in MAX PLANCK, Über irreversible Strahlungsvorgänge, Sitzungsberichte der Preußischen Akademie der Wissenschaften, Berlin, pp. 440–480, 1899. In this paper, Planck used the letter b for what nowadays is called h. Cited on page 657.

649 Bohr explained the indivisibilty of the quantum of action in his famous Como lecture. N. BOHR, *Atomtheorie und Naturbeschreibung*, Springer, Berlin, 1931. On page 16 he writes: "No more is it likely that the fundamental concepts of the classical theories will ever become superfluous for the description of physical experience. The recognition of the indivisibility of the quantum of action, and the determination of its magnitude, not only depend on an analysis of measurements based on classical concepts, but it continues to be the application of these concepts alone that makes it possible to relate the symbolism of the quantum theory to the data of experience." He also writes: "...the fundamental postulate of the indivisibility of the quantum of action is itself, from the classical point of view, an irrational element which inevitably requires us to forgo a causal mode of description and which, because of the coupling between phenomena and their observation, forces us to adopt a new mode of description designated as complementary in the sense that any given application of classical concepts precludes the simultaneous use of other classical concepts which in a different connection are equally necessary for the elucidation of the phenomena..." and "...the finite magnitude of the quantum of action prevents altogether a sharp distinction being made between a phenomenon and the agency by which it is observed, a distinction which underlies the customary concept of observation and, therefore, forms the basis of the classical ideas of motion." Other statements about the indivisibility of the quantum of action can be found in N. BOHR, *Atomic Physics and Human Knowledge*, Science Editions, New York, 1961. See also MAX JAMMER, *The Philosophy of Quantum Mechanics*, Wiley, first edition, 1974, pp. 90–91.

For some of the rare modern publications emphasizing the quantum of action see M.B. Mensky, *Physics Letters A* The action uncertainty principle and quantum gravity, **162**, p. 219, 1992, and M.B. Mensky, The action uncertainty principle in continuous quantum measurements, *Physics Letters A* **155**, pp. 229–235, 1991.

Schwinger's quantum action principle is also used in Richard F.W. Bader, *Atoms in Molecules - A Quantum Theory*, Oxford University Press, 1994.

There is a large number of general textbooks on quantum theory. The oldest textbooks are obviously German or English. The choice should be taken by the reader, considering his own preferences.

A well-known *conceptual* introduction is Jean-Marc Lévy-Leblond & Françoise Balibar, *Quantique – Rudiments*, Masson, 1997, translated into English as *Quantics*, North-Holland, 1990.

One of the most beautiful books is Julian Schwinger, *Quantum Mechanics - Symbolism of Atomic Measurements*, edited by Berthold-Georg Englert, Springer Verlag, 2001.

A modern approach with a beautiful introduction is Max Schubert & Gerhard Weber, *Quantentheorie – Grundlagen und Anwendungen*, Spektrum Akademischer Verlag, 1993.

A standard beginner's text is C. Cohen-Tannoudji, B. Diu & F. Laloë, *Mécanique quantique I et II*, Hermann, Paris, 1977. It is also available in several translations.

A good text isAsher Peres, *Quantum theory – Concepts and Methods*, Kluwer, 1995. Other textbooks regularly appear around the world. Cited on page 657.

650 Max Born, Zur Quantenmechanik der Stoßvorgänge (vorläufige Mitteilung), *Zeitschrift für Physik* **37**, pp. 863–867, 1926, Max Born, Quantenmechanik der Stoßvorgänge, *Zeitschrift für Physik* **38**, pp. 803–827, 1926. Cited on page 659.

651 See for example the papers by Jan Hilgevoord, The uncertainty principle for energy and time, *American Journal of Physics* **64**, pp. 1451–1456, 1996, and by Paul Busch, On the time–energy uncertainty reaction, part 1 & 2, *Foundations of Physics* **20**, pp. 1–43, 1990. A classic is the paper by Eugene P. Wigner, On the time–energy uncertainty reaction, in Abdus Salam & Eugene P. Wigner, editors, *Aspects of Quantum Theory*, Cambridge University Press, 1972. Cited on page 661.

652 See also the booklet by Claus Mattheck, *Warum alles kaputt geht - Form und Versagen in Natur und Technik*, Forschungszentrum Karlsruhe, 2003. Cited on page 664.

653 R. Clifton, J. Bub & H. Halvorson, Characterizing quantum theory in terms of information-theoretic constraints, http://www.arxiv.org/abs/quant-ph/0211089. Cited on page 666.

654 The quotes on motion are found in chapter VI of F. Engels, *Herrn Eugen Dührings Umwälzung der Wissenschaft*, Verlag für fremdsprachliche Literatur, Moskau, 1946. The book is commonly called *Anti–Dühring*. Cited on pages 667 and 687.

655 E.M. Brumberg & S.I. Vavilov, *Izvest. Akad. Nauk. Omen* Ser. 7, p. 919, 1933.. Cited on page 668.

656 Rodney Loudon, *The Quantum Theory of Light*, Oxford University Press, 2000. Cited on page 668.

657 F. Rieke & D.A. Baylor, Single-photon detection by rod cells of the retina, *Reviews of Modern Physics* **70**, pp. 1027–1036, 1998. They also mention that the eye usually works at photon fluxes between $10^8\,/\mu m^2 s$ (sunlight) and $10^{-2}\,/\mu m^2 s$ (starlight). The cones in the retina detect, in colour, light intensities in the uppermost 7 or 8 decades, whereas the rods detect, in black and white, the lower light intensities. Cited on page 670.

658 E. Fischbach, H. Kloor, R.A. Langel, A.T.Y. Lui & M. Peredo, New geomagnetic limit on the photon mass and on long-range forces coexisting with electromagnetism, *Physical Review Letters* **73**, pp. 514–517, 1994. Cited on page 670.

659 A.H. Compton, The scattering of X-rays as particles, *American Journal of Physics* **29**, pp. 817–820, 1961. Cited on page 670.

660 The famous paper is R. Hanbury Brown & R.Q. Twiss, *Nature* **178**, p. 1046, 1956. They got the idea to measure light in this way from their earlier work, which used the same method with radiowaves: R. Hanbury Brown & R.Q. Twiss, *Nature* **177**, p. 27, 1956, Cited on page 675.

661 *Science* 8 September 1995 Cited on page 675.

662 L. Mandel, Configuration-space photon number operators in quantum optics, *Physical Review* **144**, pp. 1071–1077, 1966. No citations.

663 A. Einstein, Über einen die Erzeugung und Umwandlung des Lichtes betreffenden heuristischen Standpunkt, *Annalen der Physik* **17**, pp. 132–184, 1905. Cited on page 676.

664 See the summary by P.W. Milonni, Answer to question 45: What (if anything) does the photoelectric effect teach us?, *American Journal of Physics* **65**, pp. 11–12, 1997. Cited on page 676.

665 For a detailed account, See Jeffrey J. Prentis, Poincaré's proof of the quantum discontinuity of nature, *American Journal of Physics* **63**, pp. 339–350, 1995. The original papers are Henri Poincaré, Sur la théorie des quanta, *Comptes Rendus de l'Académie des Sciences (Paris)* **153**, pp. 1103–1108, 1911, as well as Henri Poincaré, Sur la théorie des quanta, *Journal de Physique (Paris)* **2**, pp. 5–34, 1912. Cited on page 677.

666 J. Jacobson, G. Björk, I. Chang & Y. Yamamoto, Photonic de Broglie waves, *Physical Review Letters* **74**, pp. 4835–4838, 1995. The first measurement was published by E.J.S. Fonseca, C.H. Monken & S. de Pádua, Measurement of the de Broglie wavelength of a multiphoton wave packet, *Physical Review Letters* **82**, pp. 2868–2671, 1995. Cited on page 677.

667 For the three photon state, see M.W. Mitchell, J.S. Lundeen & A.M. Steinberg, Super-resolving phase measurements with a multiphoton entangled state, *Nature* **429**, pp. 161–164, 2004, and for the four-photon state see, in the same edition, P. Walther, J.-W. Pan, M. Aspelmeyer, R. Ursin, S. Gasparoni & A. Zeilinger, De Broglie wavelength of a non-local four-photon state, *Nature* **429**, p. 158-161, 2004. Cited on page 677.

668 For an introduction to spueezed light, see L. Mandel, Non-classical states of the electromagnetic field, *Physica Scripta T* **12**, pp. 34–42, 1986. Cited on page 677.

669 The famous quote on single photon interference is found on page 9 of P.A.M. Dirac, The Principles of Quantum Mechanics, Clarendon Press, Oxford, 1930. It is also discussed, somewhat confusely, in the otherwise informative article by H. Paul, Interference between independent photons, *Reviews of Modern Physics* **58**, pp. 209–231, 1986. Cited on page 679.

670 The original papers on coherent states are three R.J. Glauber, The quantum theory of optical coherence, *Physical Review* **130**, pp. 2529–2539, 1963, J.R. Klauder, Continuous-representation theory, I and II, *Journal of Mathematical Physics* **4**, pp. 1055–1058, 1963, and E.C.G. Sudarshan, Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams, *Physical Review Letters* **10**, p. 227, 1963. Cited on page 681.

671 R. Kidd, J. Aedini & A. Anton, Evolution of the modern photon, *American Journal of Physics* **57**, pp. 27–35, 1989, Cited on page 682.

672 WOLFGANG TITTEL, J. BRENDEL, H. ZBINDEN & N. GISIN, Violation of bell inequalities by photons more than 10 km apart, *Physical Review Letters* **81**, pp. 3563–3566, 26 October 1998. Cited on page 684.

673 N. BOHR & L. ROSENFELD, Zur Frage der Meßbarkeit der elektromagnetischen Feldgrößen, *Mat.-fys. Medd. Danske Vid. Selsk.* **12**, p. 8, 1933. The results were later published in English language as N. BOHR & L. ROSENFELD, Field and charge measurements in quantum electrodynamics, *Physical Review* **78**, pp. 794–798, 1950. Cited on page 684.

674 E.I. BUTIKOV, The rigid pendulum – an antique but evergreen physical model, *European Journal of Physics* **20**, pp. 429–441, 1999. Cited on page 685.

675 W. GERLACH & O. STERN, Der experimentelle Nachweis des magnetischen Moments des Silberatoms, *Zeitschrift für Physik* **8**, p. 110, 1921. See also the pedagogical explanation by M. HANNOUT, S. HOYT, A. KRYOWONOS & A. WIDOM, Quantum measurement and the Stern–Gerlach experiment, *American Journal of Physics* **66**, pp. 377–379, 1995. Cited on page 692.

676 J. PERRIN, Nobel Prize speech, http://www.nobel.se. H. NAGAOKA, Kinetics of a system of particles illustrating the line and the band spectrum and the phenomena of radioactivity, *Philosophical Magazine* S6, **7**, pp. 445–455, March 1904. Cited on page 697.

677 N. BOHR, On the constitution of atoms and molecules: Introduction and Part I – binding of electrons by postive nuclei, *Philosophical Magazine* **26**, pp. 1–25, 1913, On the constitution of atoms and molecules: Part II – systems containing only a single nucleus, ibid., pp. 476–502, On the constitution of atoms and molecules: Part III, ibid., pp. 857–875. Cited on page 697.

678 L. DE BROGLIE, *Comptes rendus de l'Académie des Sciences* **177**, pp. 507–510, 1923. Cited on page 688.

679 M. ARNDT, O. NAIRZ, J. VOS–ANDREAE, C. KELLER, G. VAN DER ZOUW & A. ZEILINGER, Wave–particle duality of C60 molecules, *Nature* **401**, pp. 680–682, 14 October 1999. See also the observation for tetraphenyleprophyrin and $C_{60}F_{48}$ by the same team, published as L. HACKERMÜLLER & al., Wave nature of biomolecules and fluororfullerenes, *Physical Review Letters* **91**, p. 090408, 2003.

No experiment of quantum theory has been studies as much as quantum intereference. Also the transition from interference to no interference has been epxlored, as in P. FACCHI, A. MARIANO & S. PASCAZIO, Mesoscopic interference, *Recent Developments in Physics* **3**, pp. 1–29, 2002. Cited on page 688.

680 P. CARRUTHERS & M.M. NIETO, Phase and angle variables in quantum mechanics, *Review of Modern Physics* **40**, pp. 411–440, 1968. Cited on page 691.

681 The indeterminacy relation for rotational motion is well explained by W.H. LOUISELL, Amplitude and phase uncertainty relations, *Physics Letters* **7**, p. 60, 1963. Cited on page 691.

682 S. FRANKE-ARNOLD, S.M. BARNETT, E. YAO, J. LEACH, J. COURTIAL & M. PADGETT, Uncertainty principle for angular position and angular momentum, *New Journal of Physics* **6**, p. 103, 2004. Cited on page 691.

683 ERWIN SCHRÖDINGER, Quantisierung als Eigenwertproblem I, *Annalen der Physik* **79**, pp. 361–376, 1926, and Quantisierung als Eigenwertproblem II, *Annalen der Physik* **79**, pp. 489–527, 1926. Cited on page 695.

684 C.G. GRAY, G. KARL & V.A. NOVIKOV, From Maupertius to Schrödinger. Quantization of classical variational principles, *American Journal of Physics* **67**, pp. 959–961, 1999. Cited on page 696.

685 The whole bunch of atoms behaves as one single molecule; one speaks of a Bose–Einstein condensate. The first observations, worth a Nobel prize, were by M.H. ANDERSON &

al., Observation of Bose–Einstein condensation in a dilute atomic vapour, *Science* **269**, pp. 198–201, 1995, C.C. Bradley, C.A. Sackett, J.J. Tollett & R.G. Hulet, Evidence of bose-einstein condensation in an atomic gas with attractive interactions, *Physical Review Letters* 75, pp. 1687–1690, 1995, K.B. Davis, M.-O. Mewes, M.R. Andrews, N.J. van Druten, D.S. Durfee, D.M. Kurn & W. Ketterle, Bose-Einstein Condensation in a Gas of Sodium Atoms, *Physical Review Letters* 75, pp. 3969–3973, 1995. For a simple introduction, see W. Ketterle, Experimental studies of Bose–Einstein condensation, *Physics Today* pp. 30–35, December 1999/ Cited on page 686.

686 W.M. Itano, D.J. Heinzen, J.J. Bollinger & D.J. Wineland, Quantum Zeno effect, *Physical Review A* **41**, pp. 2295–2300, 1990. M.C. Fischer, B. Gutiérrez-Medina & M.G. Raizen, Observation of the Quantum Zeno and Anti-Zeno effects in an unstable system, *Physical Review Letters* **87**, p. 040402, 2001, also http://www-arxiv.org/abs/quant-ph/0104035. Cited on page 697.

687 See P. Facchi, Z. Hradil, G. Krenn, S. Pascazio & J. Řeháček Quantum Zeno tomography, Physical Review A, **66**, p. 012110, 2002. Cited on page 697.

688 Robert H. Dicke & James P. Wittke, *Introduction to quantum theory*, Addison–Wesley, Reading, Massachusetts, 1960. See also Stephen Gasiorowicz, *Quantum Physics*, Jojn Wiley & Sons, 1974. Cited on page 690.

689 G. Papini, *Shadows of a maximal acceleration*, http://www.arxiv.org/abs/gr-qc/0211011. Cited on page 701.

690 Y. Aharonov & D. Bohm, Significance of electromagnetic potentials in the quantum theory, *Physical Review* **115**, pp. 485–491, 1959. No citations.

691 G.E. Uhlenbeck & S. Gousmit, Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons, *Naturwissenschaften* **13**, pp. 953–954, 1925. Cited on page 699.

692 L. Thomas, The motion of the spinning electron, *Nature* **117**, p. 514, 1926. Cited on page 699.

693 K. von Meyenn & E. Schucking, Wolfgang Pauli, *Physics Today* pp. 43–48, February 2001. Cited on page 699.

694 R. Pello, D. Schaerer, J. Richard, J.-F. Le Borgne & J.-P. Kneib, ISAAC,VLT observations of a lensed galaxy at z=10.0/ *Astronomy and Astrophysics* **416**, p. L35, 2004. Cited on page 704.

695 A pedagogical introduction is given by L.J. Curtis & D.G. Ellis, Use of the Einstein–Brillouin–Keller action quantization, *American journal of Physics* **72**, pp. 1521–1523, 2004. Cited on pages 704 and 706.

696 Mark P. Silverman, *And yet it moves: strange systems and subtle questions in physics*, Cambridge University Press 1993. A beautiful book from an expert of motion. Cited on pages 706, 711, and 712.

697 J. Neukammer & al., Spectroscopy of Rydberg atoms at $n \sim 500$, *Physical Review Letters* **59**, pp. 2947–2950, 1987. Cited on page 706.

698 This is explained by J.D. Hey, Mystery error in Gamow's Tompkins reappears, *Physics Today* pp. 88–89, May 2001. No citations.

699 L.L. Foldy & S.A. Wouthuysen, On the Dirac theory of spin 1/2 particles and its non-relativistic limit, *Physical Review* **78**, pp. 29–36, 1950. Cited on page 700.

700 L.L. Foldy, The electromagnetic properties of Dirac particles, *Physical Review* **83**, pp. 688–693, 1951. L.L. Foldy, The electron–neutron interaction, *Physical Review* **83**, pp.

693–696, 1951. L.L. Foldy, Electron–neutron interaction, *Review of Modern Physics* **30**, pp. 471–481, 1952. Cited on page 707.

701 John P. Costella & Bruce H.J. Mc Kellar, The Foldy–Wouthuysen transformation, *American Journal of Physics* **63**, pp. 1119–1121, 1995. Cited on page 700.

702 H. Euler & B. Kockel, *Naturwissenschaften* **23**, p. 246, 1935, H. Euler, *Annalen der Physik* **26**, p. 398, 1936, W. Heisenberg & H. Euler, Folgerung aus der Diracschen Theorie des Electrons, *Zeitschrift für Physik* **98**, pp. 714–722, 1936. Cited on page 709.

703 The g-factors for composite nuclei are explained by ... See also Hans Dehmelt, Is the electron a composite particle?, *Hyperfine Interactions* **81**, pp. 1–3, 1993. Cited on page 710.

704 J.P. Woerdman, G. Nienhuis & I. Kuščer, Is it possible to rotate an atom?, *Optics Communications* **93**, pp. 135–144, 1992. We are talking about atoms rotating around their centre of mass; atoms can of course rotate around other bodies, as discussed by M.P. Silverman, Circular birefringence of an atom in uniform rotation: the classical perspective, *American Journal of Physics* **58**, pp. 310–317, 1990. Cited on page 701.

705 For the atomic case, see P.L. Gould, G.A. Ruff & D.E. Pritchard, Diffraction of atoms by light: the near resonant Kapitza–Dirac effect, *Physical Review Letters* **56**, pp. 827–830, 1986. Many ealry experimental attempts, in particular those performed in the 1980s, for the diffraction of electrons by light were controversial; most showed only the deflection of electrons, as explained by H. Batelaan, *Contemporary Physics* **41**, p. 369, 2000. Later on, he and his group performed the newest and most spectacular experiment, demonstrating real diffraction, including interference effects; it is described in D.L. Freimund, K. Aflatooni & H. Batelaan, Observation of the Kapitza–Dirac effect, *Nature* **413**, pp. 142–143, 2001. Cited on page 711.

706 J. Schmiedmayer, M.S. Chapman, C.R. Ekstrom, T.D. Hammond, S. Wehinger & D.E. Pritchard, Index of refraction of various gases for sodium matter waves, *Physical Review Letters* **74**, p. 1043-1046, 1995. Cited on page 702.

707 The nearest to an image of a hydrogen atom is found in A. Yazdani, Watching an atom tunnel, *Nature* **409**, pp. 471–472, 2001. The experiments on Bose–Einstein condensates are also candidates for images of hydrogen atoms. The company Hitachi made a fool of itself in 1992 by claiming in a press release that its newest electron microscope could image hydrogen atoms. Cited on page 702.

708 An introduction is given by P. Pinkse & G. Rempe, Wie fängt man ein Atom mit einem Photon?, *Physikalische Blätter* **56**, pp. 49–51, 2000. Cited on page 702.

709 A.M. Wolsky, Kinetic energy, size, and the uncertainty principle, *American Journal of Physics* **42**, pp. 760–763, 1974. Cited on page 702.

710 Wolfgang Pauli, Exclusion principle and quantum mechanics, Nobel lecture, 13 December 1946, in *Nobel Lectures, Physics*, Volume 3, 1942-1962, Elsevier, 1964. Cited on page 713.

Chapter VI

Permutation of Particles

Why are we able to distinguish twins from each other? Why can we distinguish what looks alike, such as a copy from an original? In everyday life, copies always differ somewhat from originals. But think about any method that allows to distinguish objects: you will find that it runs into trouble for point-like particles. Therefore in the quantum domain something must change about our ability to distinguish objects. Let us explore the issue.

Challenge 1198 n

23. Are particles like gloves?

Some usually forgotten properties of objects are highlighted by studying a pretty combinatorial puzzle: the *glove problem*. It asks:

Ref. 711

> How many surgical gloves (for the right hand) are necessary if w doctors need to operate m patients in a hygienical way, so that nobody gets in contact with the body fluids of anybody else?

The same problem also appears in other settings. For example, it also applies to condoms, men and women – and is then called the *condom problem* – or to computers, interfaces and computer viruses. In fact, the term 'condom problem' is the term used in the books that discuss it. Obviously, the optimal number of gloves is *not* the product wm. In fact, the problem has three subcases.

Challenge 1199 e

- The simple case $m = w = 2$ already provides the most important ideas needed. Are you able to find the optimal solution and procedure?
- In the case $w = 1$ and m odd or the case $m = 1$ and w odd, the solution is $(m + 1)/2$ gloves. This is the optimal solution, as you can easily check yourself.

Challenge 1200 e

Ref. 712

- A solution with a simple procedure for all other cases is given by $\lceil 2w/3 + m/2 \rceil$ gloves, where $\lceil x \rceil$ means the smallest integer greater than or equal to x. For example, for two men and three women this gives only three gloves. (However, this formula does not always give the optimal solution; better values exist in certain subcases.)

Challenge 1201 e

Two basic properties of gloves determine the solution to the puzzle. Firstly, gloves have two sides, an interior and an exterior one. Secondly, gloves can be distinguished from each other. Do these two properties also apply to particles? We will discuss the issue of double-sidedness in the third part of the mountain ascent. In fact, the question whether particles can be turned inside out will be of great importance for their description and their motion. In the present chapter we concentrate on the second issue, namely whether objects and particles can always be distinguished. We will find that *elementary* particles do not behave like gloves in these but in an even more surprising manner. (In fact, they

Page 971

do behave like gloves in the sense that one can distinguish right-handed from left-handed ones.)

In everyday life, distinction of objects can be achieved in two ways. We are able to distinguish objects – or people – from each other because they differ in their *intrinsic properties*, such as their mass, colour, size or shape. In addition, we are also able to distinguish objects if they have the *same* intrinsic properties. Any game of billiard suggests that by following the path of each ball, we can distinguish it from the others. In short, objects with identical properties can also be distinguished using their *state*.

The state of a billiard ball is given by its position and momentum. In the case of billiard balls, the state allows distinction because the measurement error for the position of the ball is much smaller than the size of the ball itself. However, in the microscopic domain this is not the case. First of all, atoms or other microscopic particles of the same type have the same intrinsic properties. To distinguish them in collisions, we would need to keep track of their motion. But we have no chance to achieve this. Already in the nineteenth century it was shown experimentally that even nature itself is not able to do this. This result was discovered studying systems which incorporate a large number of colliding atoms of the same type: *gases*.

Page 217

The calculation of the entropy of a simple gas, made of N simple particles of mass m moving in a volume V, gives

$$S = k \ln \left[V \left(\frac{mkT}{2\pi\hbar^2} \right)^{3/2} \right]^N + \frac{3}{2} kN + k \ln \alpha \tag{539}$$

where k is the Boltzmann constant, T the temperature and ln the natural logarithm. In this formula, the pure number α is equal to 1 if the particles are distinguishable, and equal to $1/N!$ if they are not. Measuring the entropy thus allows us to determine α and therefore whether particles are distinguishable. It turns out that only the second case describes nature. This can be checked with a simple test: only in the second case does the entropy of two volumes of identical gas *add* up.* The result, often called *Gibbs' paradox*,** thus proves that the microscopic components of matter are *indistinguishable*: in a system of microscopic particles, there is no way to say which particle is which. Indistinguishability is an experimental property of nature.***

Challenge 1202 e

Ref. 713

The properties of matter would be completely different without indistinguishability. For example, we will discover that without it, knifes and swords would not cut. In addition, the soil would not carry us; we would fall right through it. To illuminate the issue

Challenge 1203 ny

* Indeed, the entropy values observed by experiment are given by the so-called Sackur–Tetrode formula

$$S = kN \ln \left[\frac{V}{N} \left(\frac{mkT}{2\pi\hbar^2} \right)^{3/2} \right] + \frac{5}{2} kN \tag{540}$$

which follows when $\alpha = 1/N!$ is inserted above.

** Josiah Willard Gibbs (1839–1903), US-American physicist who was, with Maxwell and Planck, one of the three founders of statistical mechanics and thermodynamics; he introduced the concepts of *ensemble* and of *phase*.

*** When radioactivity was discovered, people thought that it contradicted the indistinguishability of atoms, as decay seems to single out certain atoms compared to others. But quantum theory then showed that this is not the case and that atoms do remain indistinguishable.

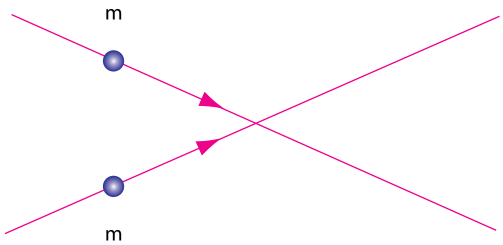


Figure 300 Identical objects with crossing paths

in more detail, we explore the next question.

Why does indistinguishability appear in nature?

Take two microscopic particles with the same mass, the same composition and the same shape, such as two atoms. Imagine that their paths cross, and that they approach each other to small distances at the crossing, as shown in Figure 300. In a gas, both the collision of atoms or a near miss are examples. Experiments show that at small distances the two particles can switch roles, without any possibility to prevent it. This is the basic process that makes it *impossible* in a gas to follow particles moving around and to determine which particle is which. This impossibility is a consequence of the quantum of action.

For a path that brings two approaching particles very close to each other, a role switch requires only a small amount of change, i.e. only a small (physical) action. However, we know that there is a smallest observable action in nature. Keeping track of each particles at small distances would require action values *smaller* than the minimal action observed in nature. The existence of a smallest action makes it thus impossible to keep track of microscopic particles when they come too near to each other. Any description of several particles must thus take into account that after a close encounter, it is impossible to say which is which.

In short, indistinguishability is a consequence of the existence of a minimal action in nature. This result leads straight away to the next question:

Can particles be counted?

In everyday life, objects can be counted because they can be distinguished. Since quantum particles cannot be distinguished, we need some care in determining how to count them. The first step is the definition of what is meant by a situation without any particle at all. This seems an easy thing to do, but later on we will encounter situations where already this step runs into difficulties. In any case, the first step is thus the *specification of the vacuum*. Any counting method requires that situations without particles be clearly separated from situations with particles.

The second step is the specification of an observable useful for determining particle number. The easiest way is to chose one of those quantum numbers which add up under composition, such as electric charge.* Counting is then performed by measuring the total

* In everyday life, the weight or mass is commonly used as observable. However, it cannot be used in the

charge and dividing by the unit charge.

This method has several advantages. First of all, it is not important whether particles are distinguishable or not; it works in all cases. Secondly, virtual particles are not counted. This is a welcome state of affairs, as we will see, because for virtual particles, i.e. for particles for which $E^2 \neq p^2c^2 + m^2c^4$, there is *no way* to define a particle number anyway.

The side effect of the counting method is that antiparticles count negatively! Also this consequence is a result of the quantum of action. We saw above that the quantum of action implies that even in vacuum, particle–antiparticle pairs are observed at sufficiently high energies. As a result, an antiparticle must count as minus one particle. In other words, any way of counting particles can produce an error due to this effect. In everyday life this limitation plays no role, as there is no antimatter around us. The issue does play a role at higher energies, however. It turns out that there is no general way to count the exact number of particles and antiparticles separately; only the sum can be defined. In short, quantum theory shows that particle counting is never perfect.

In summary, nature does provide a way to count particles even if they cannot be distinguished, though only for everyday, low energy conditions; due to the quantum of action, antiparticles count negatively, and provide a limit to the counting of particles at high energies.

What is permutation symmetry?

Since particles are countable but indistinguishable, there exists a symmetry of nature for systems composed of several identical particles. *Permutation symmetry*, also called *exchange symmetry*, is the property of nature that observations are unchanged under exchange of identical particles. Together with space-time symmetry, gauge symmetry and the not yet encountered renormalization symmetry, permutation symmetry forms one of the four pillars of quantum theory. Permutation symmetry is a property of *composed* systems, i.e. of systems made of many (identical) subsystems. Only for such systems does indistinguishability play a role.

In other words, 'indistinguishable' is not the same as 'identical'. Two particles are not the *same*; they are more like *copies* of each other. On the other hand, everyday life experience shows us that two copies can always be distinguished under close inspection, so that the term is not fully appropriate either. In the microscopic domain, particles are countable and completely indistinguishable.* Particles are perfect copies of each other.

We will discover shortly that permutation is partial rotation. Permutation symmetry thus is a symmetry under partial rotations. Can you find out why? Challenge 1205 e

Indistinguishability and symmetry

The indistinguishability of particles leads to important conclusions about the description of their state of motion. This happens because it is impossible to formulate a description of motion that includes indistinguishability right from the start. Are you able to confirm

quantum domain, except for simple cases. Can you give at least two reasons, one from special relativity and one from general relativity? Challenge 1204 n

* The word 'indistinguishable' is so long that many physicists sloppily speak of 'identical' particles nevertheless. Take care.

Challenge 1206 n this? As a consequence we describe a n-particle state with a state $\Psi_{1...i...j...n}$ which assumes that distinction is possible, as expressed by the ordered indices in the notation, and we introduce the indistinguishability afterwards. Indistinguishability means that the exchange of any two particles results in the same physical system.* Now, two quantum states have the same physical properties if they differ at most by a phase factor; indistinguishability thus requires

$$\Psi_{1...i...j...n} = e^{i\alpha}\,\Psi_{1...j...i...n} \qquad (541)$$

for some unknown angle α. Applying this expression twice, by exchanging the same couple of indices again, allows us to conclude that $e^{2i\alpha} = 1$. This implies that

$$\Psi_{1...i...j...n} = \pm\,\Psi_{1...j...i...n}\ , \qquad (542)$$

in other words, a wavefunction is either *symmetric* or *antisymmetric* under exchange of indices. Quantum theory thus predicts that particles are indistinguishable in one of two distinct ways.* Particles corresponding to symmetric wavefunctions are called *bosons*, those corresponding to antisymmetric wavefunctions are called *fermions*.**

Experiments show that the behaviour depends on the *type* of particle. Photons are bosons. On the other hand, electrons, protons and neutrons are found to be fermions. Also about half of the atoms are found to behave as bosons (at moderate energies). In fact, a composite of an *even* number of fermions (at moderate energies) – or of any number of bosons (at any energy) – turns out to be a boson; a composite of an *odd* number of fermions is (always) a fermion. For example, almost all of the known molecules are bosons (electronically speaking). Fermionic molecules are rather special and even have a special name in chemistry; they are called *radicals* and are known for their eagerness to react and to form normal bosonic molecules. Inside the human body, too many radicals can have adverse effects on health; it is well known that vitamin C is important because it is effective in reducing the number of radicals.

To which class of particles do mountains, trees, people and all other macroscopic objects belong? Challenge 1207 ny

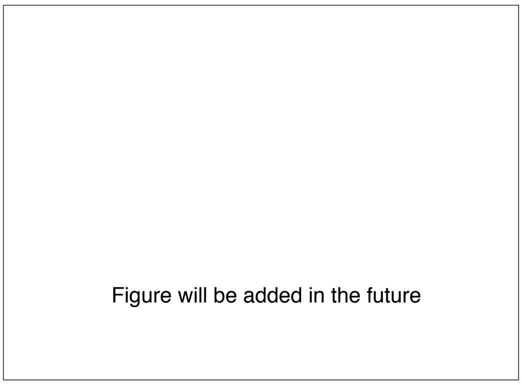


Figure 301 Two photons and interference

The behaviour of photons

A simple experiment allows to determine the behaviour of photons. Take a source that emits two photons of identical frequency and polarization at the same time, as shown in Figure 301. In the laboratory, such a source can be realized with a down-converter, a material that converts a photon of frequency 2ω into two photons of frequency ω. Both photons, after having travelled exactly the same distance, are made to enter the two sides of a beam splitter. At the two exits of the beam splitter are two detectors. Experiments Ref. 715 show that both photons are always detected together on the *same* side, and never separately on opposite sides. This result shows that photons are bosons. Fermions behave in exactly the opposite way; two fermions are always detected separately on opposite sides, never together on the same side.

The energy dependence of permutation symmetry

If experiments force us to conclude that nobody, not even nature, can distinguish any two particles of the same type, we deduce that they do not form two separate entities, but some sort of unity. Our naive, classical sense of particle as a separate entity from the rest of the world is thus an incorrect description of the phenomenon of 'particle'. Indeed, no experiment can track particles with identical intrinsic properties in such a way that they

---

* We therefore have the same situation that we encountered already several times: *an overspecification of the mathematical description*, here the explicit ordering of the indices, *implies a symmetry of this description*, which in our case is a symmetry under exchange of indices, i.e., under exchange of particles.

* This conclusion applies to three-dimensional space only. In two dimensions there are more possibilities.

** The term 'fermion' is derived from the name of the Italian physicist and Nobel Prize winner Enrico Fermi (b. 1901 Roma, d. 1954 Chicago) famous for his all-encompassing genius in theoretical and experimental physics. He mainly worked on nuclear and elementary particle physics, on spin and on statistics. For his experimental work he was called 'quantum engineer'. He is also famous for his lectures, which are still published in his own hand-writing, and his brilliant approach to physical problems. Nevertheless, his highly deserved Nobel Prize was one of the few cases in which the prize was given for a discovery which turned out to be incorrect.

'Bosons' are named after the Indian physicist Satyenra Nath Bose (b. 1894 Calcutta, d. 1974 Calcutta) Ref. 714 who first described the statistical properties of photons. The work was later expanded by Albert Einstein.

can be distinguished with certainty. This impossibility has been checked experimentally with all elementary particles, with nuclei, with atoms and with numerous molecules.

How does this fit with everyday life, i.e. with classical physics? Photons do not worry us much here. Let us focus the discussion on matter particles. We know to be able to distinguish electrons by pointing to the wire in which they flow, and we can distinguish our fridge from that of our neighbour. While the quantum of action makes distinction impossible, everyday life allows it. The simplest explanation is to imagine a microscopic particle, especially an elementary one, as a bulge, i.e. as a localized excitation of the vacuum. Figure 302 shows two such bulges representing two particles. It is evident that if particles are too near to each other, it makes no sense to distinguish them; we cannot say any more which is which.

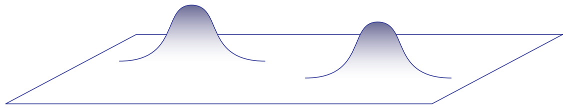


Figure 302 Particles as localized excitations

The bulge image shows that either for large distances or for high potential walls separating them, distinction of identical particles does become possible. In such situations, measurements allowing to track them independently do exist. In other words, we can specify a limit energy at which permutation symmetry of objects or particles separated

Challenge 1208 ny by a distance d becomes important. It is given by

$$E = \frac{c\,\hbar}{d} \,. \qquad (543)$$

Challenge 1209 ny Are you able to confirm the expression? For example, at everyday temperatures we *can* distinguish atoms inside a solid from each other, since the energy so calculated is much higher than the thermal energy of atoms. To have fun, you might want to determine at what energy two truly identical human twins become indistinguishable. Estimating at what energies the statistical character of trees or fridges will become apparent is then

Challenge 1210 ny straightforward.

The bulge image of particles thus purveys the idea that distinguishability exists for objects in everyday life but not for particles in the microscopic domain. To sum up, in daily life we are able to distinguish objects and thus people for two reasons: because they are made of *many* parts, and because we live in a *low energy* environment.

The energy issue immediately adds a new aspect to the discussion. How can we describe fermions and bosons in the presence of virtual particles and of antiparticles?

Indistinguishability in quantum field theory

Quantum field theory, as we will see shortly, simply puts the bulge idea of Figure 302 into mathematical language. A situation with no bulge is called *vacuum state*. Quantum field theory describes all particles of a given type as *excitations* of a single fundamental field. Particles are indistinguishable because each particle is an excitation of the same basic substrate and each excitation has the same properties. A situation with one particle is then described by a vacuum state acted upon by a *creation operator*. Adding a second particle is described by adding a second creation operator, and subtracting a particle by adding a *annihilation operator*; the latter turns out to be the adjunct of the former.

Quantum field theory then studies how these operators must behave to describe observations.* It arrives at the following conclusions:

- Fields with half-integer spin are fermions and imply (local) anticommutation.
- Fields with integer spin are bosons and imply (local) commutation.
- For all fields at spacelike separations, the commutator – respectively anticommutator – vanishes.
- Antiparticles of fermions are fermions, and antiparticles of bosons are bosons.
- Virtual particles behave like their real counterparts.

These connections are at the basis of quantum field theory. They describe how particles are identical. But why are they? Why are all electrons identical? Quantum field theory describes electrons as identical excitations of the vacuum, and as such as identical by construction. Of course, this answer is only partially satisfying. We will find a better one only in the third part of our mountain ascent.

How accurately is permutation symmetry verified?

A simple but effective experiment testing the fermion behaviour of electrons was carried out by Ramberg and Snow. They sent an electric current of 30 A through a copper wire for one month and looked for X-ray emission. They did not find any. They concluded that electrons are always in an antisymmetric state, with a symmetric component of less than

Ref. 716

$$2 \cdot 10^{-26} \tag{546}$$

of the total state. Electrons are thus fermions.

The reasoning behind this elegant experiment is the following. If electrons would not always be fermions, every now and then an electron could fall into the lowest energy level of a copper atom, leading to X-ray emission. The lack of such X-rays implies that electrons are fermions to a very high accuracy. In particular, two electrons cannot be in the same state; this is the so-called Pauli exclusion principle, our next topic.

Copies, clones and gloves

Can classical systems be indistinguishable? They can: large molecules are examples – provided they are made of exactly the same isotopes. Can *large* classical systems, made of a mole or more particles be indistinguishable? This simple question effectively asks whether a perfect copy, or (physical) *clone* of a system is possible.

It could be argued that any factory for mass-produced goods, such as one producing shirt buttons or paper clips, shows that copies are possible. But the appearance is deceiving. On a microscope there is usually some difference. Is this always the case? In 1982, the

* Whenever the relation

$$[b, b^{\dagger}] = bb^{\dagger} - b^{\dagger}b = 1 \tag{544}$$

holds between the creation operator $b^{\dagger}$ and the annihilation operator b, the operators describe a *boson*. If the operators for particle creation and annihilation anticommute

$$\{d, d^{\dagger}\} = dd^{\dagger} + d^{\dagger}d = 1 \tag{545}$$

they describe a *fermion*. The so defined bracket is called the *anticommutator bracket*.

Dutch physicist Dennis Dieks and independently, the US-American physicists Wootters Ref. 718 and Zurek, published simple proofs that quantum systems cannot be copied. This is the famous *no-cloning theorem*.

A copying machine is a machine that takes an original, reads out its properties,nd produces a copy, leaving the original unchanged. However, we know that if we extract information from an original, we have to interact with it. As a result, the system will change at least by the quantum of action. Copies and originals can never be identical.

Simply stated, if a copying machine would be able to copy originals either in state $|A\rangle$ or in state $|B\rangle$, it could not decide whether the state $|A+B\rangle$ should be copied into $|A+B\rangle$ or into $|A\rangle + |B\rangle$. This impossibility is valid for any such a machine, which is necessarily linear.* Another way to put the result is the following: if we could clone systems, we would be able to measure a variable on one system and a second variable on the copy at the same time. If we took two conjugate variables, we would be thus able to beat the indeterminacy relation. This is impossible. Copies are always imperfect.

Other researchers then explored how near to perfection a copy can be, especially in the Ref. 719 case of classical systems. To make a long story short, these investigations show that also the copying or cloning of macroscopic systems is impossible. In simple words, copying machines do not exist. Copies can always be distinguished from originals if observations are made with sufficient care. In particular, this is the case for biological clones; biological clones are identical twins born following separate pregnancies. They differ in their finger prints, iris scans, physical and emotional memories, brain structures, and in many other Challenge 1211 n aspects. (Can you specify a few more?) Biological clones, like identical twins, are not copies of each other.

The lack of quantum mechanical copying machines is disappointing. Such machines, or teleportation devices, could be fed with two different inputs, such as a lion and a goat, and produced a superposition: a chimaera. Quantum theory shows that all these imaginary beings cannot be realized.

In summary, everyday life objects such as photocopies, billiard balls or twins are always distinguishable. There are two reasons: firstly, quantum effects play no role in everyday life, so that there is no danger of unobservable exchange; secondly, perfect clones of classical systems do not exist anyway, so that there always are tiny differences between any two objects, even if they look identical at first sight. Gloves can always be distinguished.

24. Rotations and statistics – visualizing spin

We saw above that spin is the observation that matter rays, like light rays, can be polarized. Spin thus describes how particles behave under rotations, and it proves that particles are not simple spheres shrunk to points. We also saw that spin describes a fundamental difference between quantum systems and gloves: spin specifies the indistinguishability of quantum systems. Let us explore this connection in more detail.

Ref. 720 The general background for the appearance of spin was clarified by Eugene Wigner in

* The no-cloning theorem puts severe limitations on quantum computers, as computations often need copies of intermediate results. It also shows that faster-than-light communication is impossible in EPR experiments. In compensation, quantum cryptography becomes possible – at least in the laboratory. Indeed, the no-cloning theorem shows that nobody can copy a quantum message without being noticed. The specific

1939.* He started by recapitulating that any quantum mechanical particle, if elementary, must behave like an irreducible representation of the set of all viewpoint changes. This set forms the symmetry group of flat space-time, the so-called *inhomogeneous Lorentz group*. We have seen in the chapter on classical mechanics how this connection between elementarity and irreducibility arises. To be of physical relevance for quantum theory, representations have to be *unitary*. The full list of irreducible unitary representations of viewpoint changes thus provides the range of possibilities for any particle that wants to be *elementary*.

Page 177

Cataloguing the possibilities, one finds first of all that every elementary particle is described by four-momentum – no news so far – and by an internal angular momentum, the *spin*. Four-momentum results from the translation symmetry of nature, and spin from its rotation symmetry. The momentum value describes how a particle behaves under translation, i.e. under position and time shift of viewpoints. The spin value describes how an object behaves under rotations in three dimensions, i.e. under orientation change of viewpoints.* As is well known, the magnitude of four-momentum is an invariant property, given by the mass, whereas its orientation in space-time is free. Similarly, the magnitude of spin is an invariant property, and its orientation has various possibilities with respect to the direction of motion. In particular, the spin of massive particles behaves differently from that of massless particles.

For *massive* particles, the inhomogeneous Lorentz group implies that the invariant magnitude of spin is $\sqrt{J(J+1)}\,\hbar$, often simply written J. Since the value specifies the magnitude of the angular momentum, it gives the representation under rotations of a given particle type. The spin magnitude J can be any multiple of $1/2$, i.e. it can take the values $0, 1/2, 1, 3/2, 2, 5/2$, etc. Experiments show that electrons, protons and neutrons have spin 1/2, the W and Z particles spin 1 and helium atoms spin 0. In addition, the representation of spin J is $2J+1$ dimensional, meaning that the spatial orientation of the spin has $2J+1$ possible values. For electrons there are thus two possibilities; they are usually called 'up' and 'down'.

Spin thus only takes *discrete* values. This is in contrast with linear momentum, whose representations are infinite dimensional and whose possible values form a *continuous* range.

Also *massless* particles are characterized by the value of their spin. It can take the same values as in the massive case. For example, photons and gluons have spin 1. For massless particles, the representations are one-dimensional, so that massless particles are completely described by their *helicity*, defined as the projection of the spin onto the direction of motion. Massless particles can have positive or negative helicity, often also called right-handed and left-handed. There is no other freedom for the orientation of spin in the massless case.

The symmetry investigations lead to the classification of particles by their mass, their momentum and their spin. To complete the list, the remaining symmetries must be in-

ways to achieve this are the 1984 Bennett–Brassard protocol and the 1992 Ekert protocol.

* Eugene Wigner (b. 1902 Budapest, d. 1995 Princeton), Hungarian–US-American theoretical physicist, received the Nobel prize for physics in 1993. He wrote over 500 papers, many about symmetry in physics. He was also famous for being the most polite physicist in the world.

* The group of physical rotations is also called SO(3), since mathematically it is described by the group of Special Orthogonal 3 by 3 matrices.

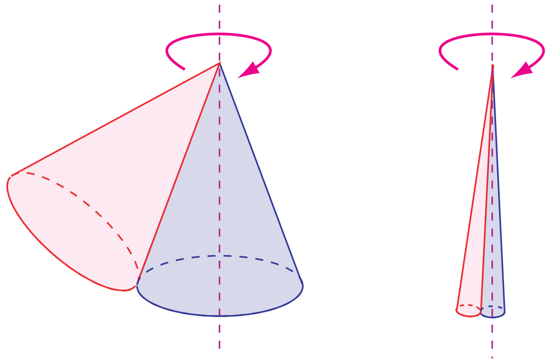


Figure 303 An argument showing why rotations by 4π are equivalent to no rotation at all

cluded. These are motion inversion parity, spatial parity and charge inversion parity. Since these symmetries are parities, each elementary particle has to be described by three additional numbers, called T, P and C, each of which can take values of either +1 or −1. Being parities, they must be *multiplied* to yield the value for a composed system.

Page 1078 A list of the values observed for all elementary particles in nature is given in Appendix C. Spin and parities together are called *quantum numbers*. As we will discover later on, additional interaction symmetries will lead to additional quantum numbers. But let us return to spin.

The main result is that spin 1/2 is a possibility in nature, even though it does not appear in everyday life. Spin 1/2 means that only a rotation of 720 degrees is equivalent to one of 0 degrees, while one of 360 degrees is not, as explained in Table 55. The mathematician Page 48 Hermann Weyl used a simple image explaining this connection.

Take two cones, touching each other at their tips as well as along a line. Hold one cone and roll the other around it, as shown in Figure 303. When the rolling cone, after a full turn around the other cone, has come back to the original position, it has rotated by some angle. If the cones are wide, the rotation angle is small. If the cones are very thin, almost like needles, the moving cone has rotated by almost 720 degrees. A rotation of 720 degrees is thus similar to one by 0 degrees. If we imagine the cone angle to vary continuously, this visualization also shows that a 720 degree rotation can be continuously deformed into a 0 degree one, whereas a 360 degree rotation cannot.

To sum up, the list of possible representations thus shows that rotations *require* the existence of spin. But why then do experiments show that all fermions have half-integer spin and that all bosons have integer spin? Why do electrons obey the Pauli exclusion principle? At first, it is not clear what the spin has to do with the statistical properties of a particle.

In fact, there are several ways to show that rotations and statistics are connected. Historically, the first proof used the details of quantum field theory and was so complicated Ref. 721 that its essential ingredients were hidden. It took quite some years to convince everybody that a simple observation about belts was the central part of the proof.

Table 55 Particle spin as representation of the rotation group

Spin $[\hbar]$	System unchanged after rotation by	Massive examples elementary	composite	Massless examples elementary
0	any angle	none [a,b]	mesons, nuclei, atoms	none [b]
1/2	2 turns	e, μ, τ, q	nuclei, atoms, molecules	ν_e, ν_μ, ν_τ
1	1 turn	W, Z	mesons, nuclei, atoms, molecules, toasters	g, γ
3/2	2/3 turn	none [b]	baryons, nuclei, atoms	none [b]
2	1/2 turn	none	nuclei	'graviton' [c]
5/2	2/5 turn	none	nuclei	none
3	1/3 turn	none	nuclei [d]	none
etc.[d]	etc.[d]	etc.[d]	etc.[d]	etc.[d]

a. Whether the Higgs particle is elementary or not is still unknown.
b. Supersymmetry predicts particles in these and other boxes.
c. The graviton has not yet been observed.
d. Nuclei exist with spins values up to at least 101/2 and 51 (in units of $\hbar$). Ref. 722

The belt trick

Ref. 723 The well-known *belt trick* was often used by Dirac to explain the features of spin 1/2. Taking Figure 302, which models particles as indistinguishable excitations, it is not difficult to imagine a sort of sheet connecting them, similar to a belt connecting the two parts of the buckle, as shown in Figure 304. If one end of the belt is rotated by 2π along any axis, a twist is inserted into the belt. If the end is rotated for another 2π, bringing the total to 4π, the ensuing double twist can easily be undone without moving or rotating the ends.
Challenge 1212 e You need to experience this yourself in order to believe it.

In addition, if you take the two ends and simply *swap* positions, a twist is introduced into the belt. Again, a second swap will undo the twist.

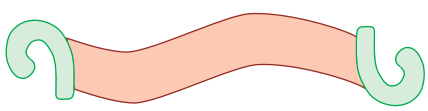


Figure 304 A belt visualizing two spin 1/2 particles

In other words, if we take each end to represent a particle and a twist to mean a factor -1, the belt exactly describes the phase behaviour of spin 1/2 wavefunctions under exchange and under rotations. In particular, we see that spin and exchange behaviour are related.

The human body has such a belt built in: the *arm*. Just take your hand, put an object on it for clarity and turn the hand and object by 2π by twisting the arm. After a second

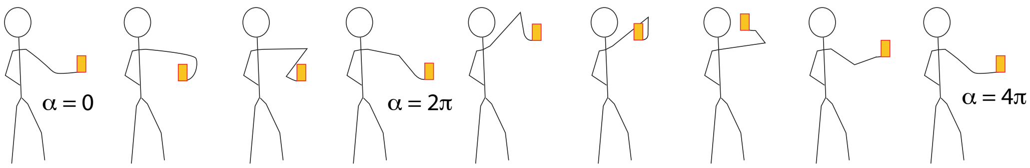


Figure 305 The human arm as spin 1/2 model

rotation the whole system will be untangled again.

The trick is even more impressive when many tails are used. In fact, there are two ways to do this. One way is connect two buckles with *many* bands or threads, like in Figure 306. Both a rotation by 2π of one end or an exchange of the two ends produces quite a tangle; nevertheless, in both cases a second operation leads back to the original situation.

There is a second, even more interesting way to show the connection between rotation and exchange. Just glue any number of threads or bands, say half a meter long, to two asymmetric objects. Like the arm of a human being, the bands are supposed to go to infinity and be attached there. If any of the objects, which represent the particles, is rotated by 2π, twists appear in its strings. If the object is rotated by an additional turn, to a total of 4π, as shown in Figure 307, all twists and tangles can be made to disappear, without moving or turning the object. You really have to experience this in order to believe it. And the trick really works with *any* number of bands glued to the object.

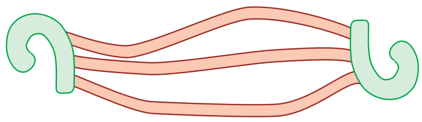


Figure 306 Another model for two spin 1/2 particles

Even more astonishing is the other half of the experiment. Take *two* particles as shown in the left of Figure 307. If you exchange the positions of two such spin 1/2 particles, always keeping the ends at infinity fixed, a tangled mess is created. But incredibly, if you exchange the objects a second time, everything untangles neatly, independently of the number of
Challenge 1213 e attached strings. You might want to test yourself that the behaviour is still the same with sets of three or more particles.

All these observations together form the spin statistics theorem for spin 1/2 particles: *spin and exchange behaviour are related*. Indeed, these almost 'experimental' arguments
Ref. 724 can be put into exact mathematical language by studying the behaviour of the configuration space of particles. These investigations result in the following statements:

▷ *Objects of spin 1/2 are fermions.**

▷ *Exchange and rotation of spin 1/2 particles are similar processes.*

Note that all these arguments require three dimensions, because there are no tangles (or knots) in fewer dimensions.** And indeed, spin exists only in three or more spatial dimensions.

* A mathematical observable behaving like a spin 1/2 particle is neither a vector nor a tensor, as you may
Challenge 1214 e want to check. An additional concept is necessary; such an observable is called a *spinor*. We will introduce
Page 707 it later on.

** Of course, knots and tangles do exist in higher dimensions. Instead of considering knotted one-dimensional lines, one can consider knotted planes or knotted higher-dimensional hyperplanes. For ex-

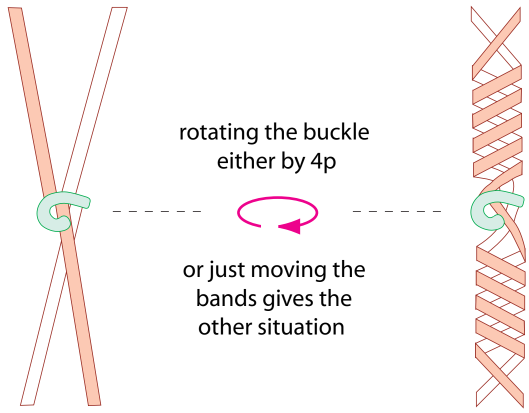


Figure 307 The extended belt trick, modelling a spin 1/2 particle: the two situations can be transformed into each other either by rotating the central object by 4π or by keeping the central object fixed and moving the bands around it

The Pauli exclusion principle and the hardness of matter

Why are we able to knock on a door? Why can stones not fly through tree trunks? How does the mountain we are walking on carry us? In classical physics, we avoided this issue, by taking solidity as a defining property of matter. But doing so, we cheated: we have seen that matter consists mainly of empty space, so that we have to study the issue without any sneaky way out. The answer is not clear: penetration is made impossible by Pauli's exclusion principle between the electrons inside atoms.

Why do electrons and other fermions obey the Pauli exclusion principle? The answer can be given with a beautifully simple argument. We know that exchanging two fermions produces a minus sign. Imagine these two fermions being, as a classical physicist would say, located at the same spot, or as a quantum physicist would say, in the same state. If that could be possible, an exchange would change nothing in the system. But an exchange of fermions must produce a minus sign for the total state. Both possibilities – no change at all as well as a minus sign – cannot be realized at the same time. There is only one way out: two fermions must avoid to ever be in the same state.

Ref. 725

The exclusion principle is the reason that two pieces of matter in everyday life cannot penetrate each other, but have to repel each other. For example, bells only work because of the exclusion principle. Bells would not work if the colliding pieces that produce the sound would interpenetrate. But in any example of two interpenetrating pieces the electrons in the atoms would have to be in similar states. This is forbidden. For the same reason we do not fall through the floor, even though gravity pulls us down, but remain on the surface. In other words, the exclusion principle implies that matter cannot be compressed indefinitely, as at a certain stage an effective Pauli pressure appears, so that a compression limit ensues. For this reason for example, planets or neutron stars do not collapse under their own gravity.

The exclusion principle also answers the question about how many angels can dance on the top of a pin. (Note that angels must be made of fermions, as you might want to

ample, deformable planes can be knotted in four dimensions and deformable 3-spaces in five dimensions.

Challenge 1215 n deduce from the information known about them.) Both theory and experiment confirm Ref. 717 the answer already given by Thomas Aquinas in the middle ages: only one. The fermion exclusion principle could also be called 'angel exclusion principle'. To stay in the topic, the principle also shows that *ghosts* cannot be objects, as ghosts are supposed to be able to traverse walls.

Whatever the interpretation, the exclusion principle keeps things in shape; without it, there would be no three-dimensional objects. Only the exclusion principle keeps the cloudy atoms of nature from merging, holding them apart. Shapes are a direct consequence of the exclusion principle. As a result, when we knock on a table or on a door, we show that both objects are made of fermions.

Since permutation properties and spin properties of fermions are so well described by the belt model, we could be led to the conclusion that these properties might really be consequence of such belt-like connections between particles and the outside world. Maybe for some reason we only observe the belt buckles, not the belts themselves. In the third part of this walk we will discover whether this idea is correct.

So far, we have only considered spin 1/2 particles. We will not talk much about systems with odd spin of higher value, such as 3/2 or 5/2. Such systems can be seen as being Challenge 1216 ny composed of spin 1/2 entities. Can you confirm this?

We did not talk about lower spins than 1/2 either. A famous theorem states that a positive spin value below 1/2 is impossible, because the largest angle that can be measured in three dimensions is 4π. There is no way to measure a larger angle;* The quantum of action makes this impossible. Thus there cannot be any spin value between 0 and 1/2.

Integer spin

Under rotations, integer spin particles behave differently from half-integer particles. Integer spin particles do not show the strange sign changes under rotations by 2π. In the belt imagery, integer spin particles need no attached strings. The spin 0 particle obviously corresponds to a sphere. Models for other spin values are shown in Figure 308. Exploring their properties in the same way as above, we arrive at the so-called *spin-statistics theorem*:

▷ *Exchange and rotation of objects are similar processes.*

▷ *Objects of half-integer spin are fermions. They obey the Pauli exclusion principle.*

▷ *Objects of integer spin are bosons.*

Challenge 1217 ny You might prove by yourself that this suffices to show that

▷ *Composites of bosons, as well as composites of an even number of fermions (at low energy), are bosons; composites of an uneven number of fermions are fermions.***

* This is possible in two dimensions though.

** This sentence implies that spin 1 and higher can also be achieved *with* tails; can you find such a repres- Challenge 1218 ny entation?

Note that composite fermions can be bosons only up to that energy at which the composition breaks down. Otherwise, by packing fermions into bosons, we could have fermions in the same state.

These connections express basic characteristics of the three-dimensional world in which we live.

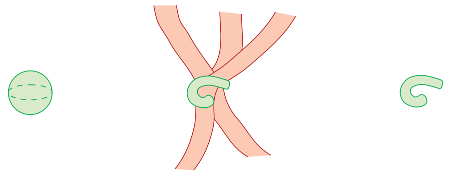


Figure 308 Some visualizations of spin representations

Is spin a rotation about an axis?

The spin of a particle behaves experimentally like an intrinsic angular momentum, adds up like angular momentum, is conserved as part of angular momentum, is described like angular momentum and has a name synonymous with angular momentum. Despite all this, for many decades a strange myth was spread in physics courses and textbooks around the world, namely that spin 1/2 is *not* a rotation about an axis. The myth maintains that any rotating object must have integer spin. Since half integer spin is not possible in classical physics, it is argued that such spin is not due to rotation. It is time to finish with this example of muddled thinking.

Electrons do have spin 1/2 and are charged. Electrons and all other charged particles with spin 1/2 do have a magnetic moment.* A magnetic moment is expected for any rotating charge. In other words, spin 1/2 does behave like rotation. However, assuming that a particle consists of a continuous charge distribution in rotational motion gives the wrong value for the magnetic moment. In the early days of the twentieth century, when physicists were still thinking in classical terms, they concluded that spin 1/2 particles thus cannot be rotating. This myth has survived through many textbooks. The correct deduction is that the assumption of continuous charge distribution is wrong. Indeed, charge is quantized; nobody today expects that elementary charge is continuously spread over space, as that would contradict its quantization.

Let us remember what rotation is. Both the belt trick for spin 1/2 as well as the integer spin case remind us: a *rotation* of one body around another is a fraction or a multiple of an exchange. What we call a rotating body in everyday life is a body continuously exchanging the positions of its parts. Rotation and exchange are the same.

Above we found that spin is exchange behaviour. Since rotation is exchange and spin is exchange, it follows that spin *is* rotation. Since we deduced, like Wigner, spin from rotation invariance, this consequence is not a surprise.

The belt model of a spin 1/2 particle tells us that such a particle can rotate continuously
Challenge 1220 ny without any hindrance. In short, we are allowed to maintain that spin is rotation about an axis, without any contradiction to observations, even for spin 1/2. The belt model helps to keep two things in mind: we must assume that in the belt model only the buckles can
Ref. 726 be observed and do interact, not the belts, and we must assume that elementary charge is pointlike and cannot be distributed.**

Challenge 1219 ny * This can easily be measured in a an experiment; however, not one of the Stern–Gerlach type. Why?

** Obviously, the detailed structure of the electron still remains unclear at this point. Any angular momentum S is given classically by $S = \Theta\omega$; however, neither the moment of inertia Θ, connected to the rotation radius and electron mass, nor the angular velocity ω are known at this point. We have to wait quite a while, until the third part of our adventure, to find out more.

Why is fencing with laser beams impossible?

When a sword is approaching dangerously, we can stop it with a second sword. Many old movies use such scenes. When a laser beam is approaching, it is impossible to fend it off with a second beam, despite all science fiction movies showing so. Banging two laser beams against each other is impossible.

The above discussion shows why. The electrons in the swords are fermions and obey the Pauli exclusion principle. Fermions make matter impenetrable. On the other hand, photons are bosons. Bosons can be in the same state; they allow penetration. Matter is impenetrable because at the fundamental level it is composed of fermions. Radiation is composed of bosons. The distinction between fermions and bosons thus explains why objects can be touched while images cannot. In the first part of our mountain ascent we Page 70 started by noting this difference; now we know its origin.

Rotation requires antiparticles

The connection between rotation and antiparticles may be the most astonishing conclusion from the experiments showing the existence of spin. So far, we have seen that rotation requires the existence of spin, that spin appears when relativity is introduced into Page 287 quantum theory, and that relativity requires antimatter. Taking these three statements together, the conclusion of the title is not surprising any more. Interestingly, there is a simple argument making the same point directly, without any help of quantum theory, when the belt model is extended from space alone to full *space-time*.

To learn how to think in space-time, let us take a particle spin 1, i.e. a particle looking like a detached belt buckle in three dimensions. When moving in a 2+1 dimensional space-time, it is described by a ribbon. Challenge 1221 ny Playing around with ribbons in space-time, instead of belts in space, provides many interesting conclusions. For example, Figure 309 shows that wrapping a rubber ribbon around the fingers can show that a rotation of a body by 2π in presence of a second one is the same as exchanging the positions of the two bodies.* Both sides of the hand transform the same initial condition, at one border of the hand, to the same final condition at the other border. We have thus successfully extended a known result from space to space-time. Interestingly, we can also find a smooth sequence of steps realizing this equivalence.

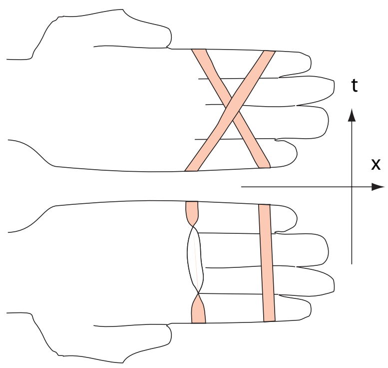


Figure 309 Equivalence of exchange and rotation in space-time

When particles in space-time are described as ribbons, Figure 310 shows the intermediate steps allowing to identify a rotation with an exchange. The sequence requires the use of a particle–antiparticle pair. Without antiparticles, the equivalence of rotation and exchange would not hold in space-time. Rotation in space-time indeed requires antiparticles.

* Obviously, the next step would be to check the full spin 1/2 model of Figure 307 in four-dimensional space-time. But this is not an easy task; there is no generally accepted solution yet. Challenge 1222 ny

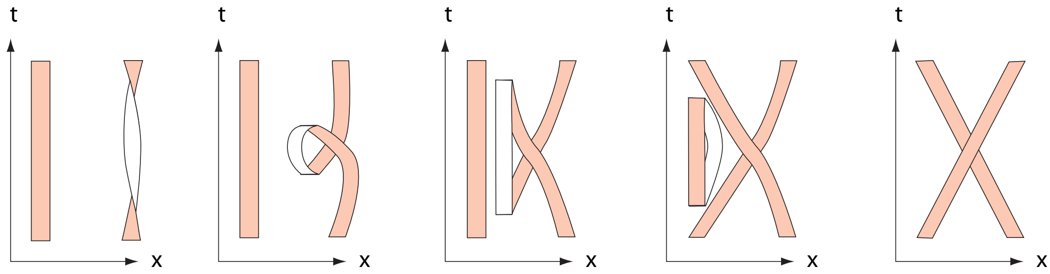


Figure 310 Belts in space-time: rotation and antiparticles

Limits and open questions of quantum statistics

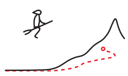
The topic of statistics is an important research field in theoretical and experimental physics. In particular, researchers have searched and still are searching for generalizations of the possible exchange behaviours of particles.

In two spatial dimensions, the result of an exchange of the wavefunction is not described by a sign, but by a continuous phase. Two-dimensional objects behaving in this way, called *anyons* because they can have 'any' spin, have experimental importance, since in many experiments in solid state physics the set-up is effectively two-dimensional. The fractional quantum Hall effect, perhaps the most interesting discovery of modern experimental physics, has pushed anyons onto the stage of modern research. Page 791

Other theorists generalized the concept of fermions in other ways, introducing parafermions, parabosons, plektons and other hypothetical concepts. O.W. Greenberg has spent most of his professional life on this issue. His conclusion is that in 3 + 1 space-time dimensions, only fermions and bosons exist. (Can you show that this implies that the ghosts appearing in scottish tales do not exist?) Ref. 727 Challenge 1223 n

From a different viewpoint, the above belt model invites to study the behaviour of braids and knots. (In mathematics, a *braid* is a knot extending to infinity.) This fascinating part of mathematical physics has become important with the advent of string theory, which states that particles, especially at high energies, are not point-like, but extended entities.

Still another generalization of statistical behaviour at high energies is the concept of quantum group, which we will encounter later on. In all of these cases, the quest is to understand what happens to permutation symmetry in a unified theory of nature. A glimpse of the difficulties appears already above: how can Figures 302, 307 and 310 be reconciled? We will settle this issue in the third part of our mountain ascent.



Bibliography

711 See the paper by Martin Gardner, Science fiction puzzle tales, Clarkson Potter, **67**, pp. 104–105, 1981, or his book *Aha! Insight*, Scientific American & W.H. Freeman, New York,

1978. The rabbit story is from A. Hajnal & P. Lovász, An algorithm to prevent the propagation of certain diseases at minimum cost, in *Interfaces Between Computer Science and Operations Research*, edited by J.K. Lenstra, A.H.G. Rinnooy Kan & P. Van Emde Boas, Mathematisch Centrum, Amsterdam 1978, whereas the computer euphemism is used by A. Orlitzky & L. Shepp, On curbing virus propagation, Technical memorandum, Bell Labs 1989. Cited on page 719.

712 A complete discussion of the problem can be found in chapter 10 of Ilan Vardi, *Computational Recreations in Mathematica*, Addison Wesley, Redwood City, California, 1991. Cited on page 719.

713 On Gibbs' paradox, see your favourite text on thermodynamics or statistical mechanics. See also W.H. Zurek, Algorithmic randomness and physical entropy, *Physical Review A* **40**, pp. 4731–4751, 1989. Zurek shows that the Sackur–Tetrode formula can be derived from algorithmic entropy considerations. Cited on page 720.

714 S.N. Bose, Plancks Gesetz und Lichtquantenhypothese, *Zeitschrift für Physik* **26**, pp. 178–181, 1924. The theory was then expanded in A. Einstein, Quantentheorie des einatomigen idealen Gases, *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin* **22**, pp. 261–267, 1924, A. Einstein, Quantentheorie des einatomigen idealen Gases. Zweite Abhandlung, *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin* **23**, pp. 3–14, 1925, A. Einstein, Zur Quantentheorie des idealen Gases, *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin* **23**, pp. 18–25, 1925. Cited on page 724.

715 C.K. Hong, Z.Y. Ou & L. Mandel, Measurement of subpicosecond time intervals between two photons by interference, *Physical Review Letters* **59**, pp. 2044–2046, 1987. Cited on page 724.

716 The experiment is described in E. Ramberg & G.A. Snow, Experimental limit on a small violation of the Pauli principle, *Physics Letters B* **238**, pp. 438–441, 1990. Other experimental tests are reviewed in O.W. Greenberg, Particles with small violations of Fermi or Bose statistics, *Physical Review D* **43**, pp. 4111–4120, 1991. Cited on page 726.

717 The issue is treated in his *Summa Theologica*, in question 52 of the first part. The complete text, several thousand pages, can be found on the http://www.newadvent.org website. Cited on page 733.

718 The original no-cloning theorem is by D. Dieks, Communication by EPR devices, *Physics Letters A* **92**, pp. 271–272, 1982, and by W.K. Wootters & W.H. Zurek, A single quantum cannot be cloned, *Nature* **299**, pp. 802–803, 1982. For a discussion of photon and multiparticle cloning, see N. Gisin & S. Massar, Optimal quantum cloning machines, *Physics Review Letters* **79**, pp. 2153–2156, 1997. The whole topic has been presented in detail by V. Buzek & M. Hillery, Quantum cloning, *Physics World* **14**, pp. 25–29, November 2001. Cited on page 727.

719 The most recent experimental and theoretical results on physical cloning are described in A. Lamas-Linares, C. Simon, J.C. Howell & D. Bouwmeester, Experimental quantum cloning of single photons, *Science* **296**, pp. 712 – 714, 2002, D. Collins & S. Popescu, A classical analogue of entanglement, preprint http://www.arxiv.org/abs/quant-ph/0107082, 2001, and A. Daffertshofer, A.R. Plastino & A. Plastino, Classical no-cloning theorem, *Physical Review Letters* **88**, p. 210601, 2002. Cited on page 727.

720 E. Wigner, On unitary representations of the inhomogeneous Lorentz group, *Annals of Mathematics* **40**, pp. 149–204, 1939. This famous paper summarises the work which later brought him the Nobel Prize for physics. Cited on page 727.

721 WOLFGANG PAULI, The Connection Between Spin and Statistics, *Physical Review* **58**, pp. 716– 722, 1940. Cited on page 729.

722 For a full list of isotopes, see R.B. FIRESTONE, *Table of Isotopes, Eighth Edition, 1999 Update*, with CDROM, John Wiley & Sons, 1999. Cited on page 730.

723 This famous explanation is given, *e.g.*, on page 1148 in C.W. MISNER, K.S. THORNE & J.A. WHEELER, *Gravitation*, Freeman, 1973. It is called the *scissor trick* on page 43 of volume 1 of R. PENROSE & W. RINDLER, *Spinors and Spacetime*, 1984. It is also cited and discussed by R. GOULD, Answer to question #7, *American Journal of Physics* **63**, p. 109, 1995. Cited on page 730.

724 M.V. BERRY & J.M. ROBBINS, Indistinguishability for quantum particles: spin, statistics and the geometric phase, *Proceedings of the Royal Society in London A* **453**, pp. 1771–1790, 1997. See also the comments to this result by J. TWAMLEY, Statistics given a spin, *Nature* **389**, pp. 127–128, 11 September 1997. Their newer results are M.V. BERRY & J.M. ROBBINS, Quantum indistinguishability: alternative constructions of the transported basis, *Journal of Physics A (Letters)* **33**, pp. L207–L214, 2000, and M.V. BERRY & J.M. ROBBINS, in *Spin-Statistics*, eds. R. HILBORN & G. TINO, (American Institute of Physics), 2000, pp. 3–15. See also Michael Berry's home page on the http://www.phy.bris.ac.uk/staff/berry_mv.html website. Cited on page 731.

725 R.W. HARTUNG, Pauli principle in Euclidean geometry, *American Journal of Physics* **47**, pp. 900–910, 1979. Cited on page 732.

726 The point that spin can be seen as a rotation was already made by F.J. BELINFANTE, On the spin angular momentum of mesons, *Physica* **6**, p. 887, 1939, and taken up again by HANS C. OHANIAN, What is spin?, *American Journal of Physics* **54**, pp. 500–505, 1986. See also E. DURAN & A. ERSCHOW, *Physikalische Zeitschrift der Sowjetunion* **12**, p. 466, 1937. Cited on page 734.

727 Generalizations of bosons and fermions are reviewed in the (serious!) paper by O.W. GREENBERG, D.M. GREENBERGER & T.V. GREENBERGEST, (Para)bosons, (para)fermions, quons and other beasts in the menagerie of particle statistics, at http://www.arxiv.org/abs/hep-th/9306225. A newer summary is O.W. GREENBERG, Theories of violation of statistics, electronic preprint available at http://www.arxiv.org/abs/hep-th/0007054, 2000. Cited on page 736.

Chapter VII

Details of Quantum Theory and Electromagnetism

25. Superpositions and probabilities – quantum theory without ideology

Ref. 728

> The fact that an adequate philosophical presentation has been so long delayed is no doubt caused by the fact that Niels Bohr brainwashed a whole generation of theorists into thinking that the job was done fifty years ago.
>
> Murray Gell-Mann

Why is this famous physical issue arousing such strong emotions? In particular, who is brainwashed, Gell-Mann, the discoverer of the quarks, or most of the other physicists working on quantum theory who follow Niels Bohr's* opinion?

In the twentieth century, quantum mechanics has thrown many in disarray. Indeed, it radically changed the two most basic concepts of classical physics: state and system. The *state* is not described any more by the specific values taken by position and momentum, but by the specific wavefunction 'taken' by the position and momentum operators.** In addition, in classical physics a *system* was described as a set of permanent aspects of nature; permanence was defined as negligible interaction with the environment. Quantum mechanics shows that this definition has to be modified as well.

In order to clarify the issues, we take a short walk around the strangest aspects of quantum theory. The section is essential if we want to avoid getting lost on our way to the top of Motion Mountain, as happened to quite a number of people since quantum theory appeared.

Why are people either dead or alive?

The evolution equation of quantum mechanics is linear in the wavefunction; thus we can imagine and try to construct systems where the state ψ is a superposition of two radically distinct situations, such as those of a dead and of a living cat. This famous fictional animal is called *Schrödinger's cat* after the originator of the example. Is it possible to produce it?

* Niels Bohr (b. 1885 Copenhagen, d. 1962 Copenhagen) made Copenhagen University into one of the centres of quantum theory, overshadowing Göttingen. He developed the description of the atom with quantum theory, for which he received the 1922 Nobel prize in physics. He had to flee Denmark in 1943 after the German invasion, because of his Jewish background, but returned there after the war.

** It is equivalent, but maybe conceptually clearer, to say that the state is described by a complete set of commuting operators. In fact, the discussion is somewhat simplified in the Heisenberg picture. However, here we study the issue in the Schrödinger picture, using wavefunctions.

How would it evolve in time? We can ask the same questions about a superposition of a state where a car is inside a closed garage with a state where the car is outside.

Such strange situations are not usually observed in everyday life. The reason for this rareness is an important aspect of what is often called the 'interpretation' of quantum mechanics. In principle, such strange situations are possible, and the superposition of macroscopically distinct states has actually been observed in a few cases, though not for cats, people or cars. To get an idea of the constraints, let us specify the situation in more detail.* The object of discussion are linear superpositions of the type $\psi = a\psi_a + b\psi_b$, where ψ_a and ψ_b are macroscopically distinct states of the system under discussion, and where a and b are some complex coefficients. States are called *macroscopically distinct* when each state corresponds to a different macroscopic situation, i.e. when the two states can be distinguished using the concepts or measurement methods of classical physics. In particular, this means that the physical action necessary to transform one state into the other must be much larger than $\hbar$. For example, two different positions of any body composed of a large number of molecules are macroscopically distinct.

A 'strange' situation is thus a superposition of macroscopic distinct states. Let us work out the essence of macroscopic superpositions more clearly. Given two macroscopically distinct states ψ_a and ψ_b, a superposition of the type $\psi = a\psi_a + b\psi_b$ is called a *pure state*. Since the states ψ_a and ψ_b can interfere, one also talks about a *(phase) coherent superposition*. In the case of a superposition of macroscopically distinct states, the scalar product $\psi_a^\dagger \psi_b$ is obviously vanishing. In case of a coherent superposition, the coefficient product $a^* b$ is different from zero. This fact can also be expressed with the help of the *density matrix* ρ of the system, defined as $\rho = \psi \otimes \psi^\dagger$. In the present case it is given by

no figure yet

Figure 311 Artist's impression of a macroscopic superposition

$$\rho_{\text{pure}} = \psi \otimes \psi^\dagger = |a|^2 \psi_a \otimes \psi_a^\dagger + |b|^2 \psi_b \otimes \psi_b^\dagger + a\, b^* \psi_a \otimes \psi_b^\dagger + a^*\, b\, \psi_b \otimes \psi_a^\dagger$$
$$= (\psi_a, \psi_b) \begin{pmatrix} |a|^2 & a\, b^* \\ a^*\, b & |b|^2 \end{pmatrix} \begin{pmatrix} \psi_a^\dagger \\ \psi_b^\dagger \end{pmatrix} . \quad (547)$$

We can then say that whenever the system is in a pure state, its density matrix, or *density functional*, contains off-diagonal terms of the same order of magnitude as the diagonal ones.** Such a density matrix corresponds to the above-mentioned strange situations that

* Most what can be said about this topic has been said by two people: John von Neumann, who in the
Ref. 729 nineteen-thirties stressed the differences between evolution and decoherence, and by Hans Dieter Zeh, who
Ref. 730 in the nineteen seventies stressed the importance of baths and the environment in decoherence.

** Using the density matrix, we can rewrite the evolution equation of a quantum system:

$$\dot{\psi} = -iH\psi \quad \text{becomes} \quad \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\frac{i}{\hbar}[H, \rho] . \quad (548)$$

Both are completely equivalent. (The new expression is sometimes also called the *von Neumann equation*.) We won't actually do any calculations here. The expressions are given so that you recognize them when you we do not observe in daily life.

We now have a look at the opposite situation. In contrast to the case just mentioned, a density matrix for macroscopic distinct states with *vanishing* off-diagonal elements, such as the two state example

$$\begin{aligned} \rho &= |a|^2 \psi_a \otimes \psi_a^\dagger + |b|^2 \psi_b \otimes \psi_b^\dagger \\ &= (\psi_a, \psi_b) \begin{pmatrix} |a|^2 & 0 \\ 0 & |b|^2 \end{pmatrix} \begin{pmatrix} \psi_a^\dagger \\ \psi_b^\dagger \end{pmatrix} \end{aligned} \tag{549}$$

describes a system which possesses *no* phase coherence at all. (Here, $\otimes$ denotes the non-commutative dyadic product or tensor product which produces a tensor or matrix starting from two vectors.) Such a diagonal density matrix cannot be that of a pure state; it describes a system which is in the state ψ_a with probability $|a|^2$ and which is in the state ψ_b with probability $|b|^2$. Such a system is said to be in a *mixed state*, because its state is *not known*, or equivalently, to be in a *(phase) incoherent superposition*, because interference effects cannot be observed in such a situation. A system described by a mixed state is always *either* in the state ψ_a *or* in the state ψ_b. In other words, a diagonal density matrix for macroscopically distinct states is not in contrast, but in agreement with everyday experience. In the picture of density matrices, the non-diagonal elements contain the difference between normal, i.e. incoherent, and unusual, i.e. coherent, superpositions.

The experimental situation is clear: for macroscopically distinct states, only diagonal density matrices are observed. Any system in a coherent macroscopic superposition somehow loses its off-diagonal matrix elements. How does this process of *decoherence** take place? The density matrix itself shows the way.

Indeed, the density matrix for a large system is used, in thermodynamics, for the

Ref. 731

definition of its entropy and of all its other thermodynamic quantities. These studies show

Challenge 1224 ny

that

$$S = -k \operatorname{tr}(\rho \ln \rho) \tag{550}$$

where tr denotes the *trace*, i.e. the sum of all diagonal elements. We also remind ourselves that a system with a large and constant entropy is called a *bath*. In simple physical terms, a bath is a system to which we can ascribe a temperature. More precisely, a *(physical) bath*, or *(thermodynamic) reservoir*, is any large system for which the concept of *equilibrium* can be defined. Experiments show that in practice, this is equivalent to the condition that a bath consists of many interacting subsystems. For this reason, all macroscopic quantities describing the state of a bath show small, irregular *fluctuations*, a fact that will be of central importance shortly.

Obviously, an everyday bath is also a bath in the physical sense: a thermodynamic bath is similar to an extremely large warm water bath, one for which the temperature does not change even if one adds some cold or warm water to it. Examples of physical baths are an intense magnetic field, a large amount of gas, or a large solid. (The meanings of 'intense' and 'large' of course depend on the system under study.) The physical concept of bath is thus an abstraction and a generalization of the everyday concept of bath.

encounter them elsewhere.

* In many settings, decoherence is called *disentanglement*, as we will see below.

Challenge 1225 n It is easy to see from the definition (550) of entropy that the loss of off-diagonal elements corresponds to an increase in entropy. And it is known that any increase in entropy of a reversible system, such as the quantum mechanical system in question, is due to an interaction with a bath.

Where is the bath interacting with the system? It obviously must be outside the system one is talking about, i.e. in its *environment*. Indeed, we know experimentally that any environment is large and characterized by a temperature. Some examples are listed in Table 56. Any environment therefore contains a bath. We can even go further: for every experimental situation, there is a bath *interacting* with the system. Indeed, every system which can be observed is not isolated, as it obviously interacts at least with the observer; and every observer by definition contains a bath, as we will show in more detail shortly. Usually however, the most important baths we have to take into consideration are the atmosphere around a system, the radiation attaining the system or, if the system itself is large enough to have a temperature, those degrees of freedom of the system which are not involved in the superposition under investigation.

If every system is in contact with baths, every density matrix of a macroscopic superposition lose their diagonal elements. At first sight, this direction of thought is not convincing. The interactions of a system with its environment can be made extremely small by using clever experimental set-ups; that would imply that the time for decoherence can be made extremely large. Thus we need to check how much time a superposition of states needs to decohere. It turns out that there are two standard ways to estimate the *decoherence time:* either by modelling the bath as large number of colliding particles, or by modelling it as a continuous field.

Challenge 1226 ny If the bath is described as a set of particles randomly hitting the microscopic system, it is best characterized by the effective wavelength λ_{eff} of the particles and by the average interval t_{hit} between two hits. A straightforward calculation shows that the decoherence time t_d is in any case smaller than this time interval, so that

$$t_d \leqslant t_{\text{hit}} = \frac{1}{\varphi\sigma} , \qquad (551)$$

where φ is the flux of particles and σ the cross section for the hit.* Typical values are given in Table 56. We easily note that for macroscopic objects, decoherence times are extremely short. Scattering leads to fast decoherence. However, for atoms or smaller systems, the situation is different, as expected.

* The decoherence time is derived by studying the evolution of the density matrix $\rho(x, x')$ of objects localized at two points x and x'. One finds that the off-diagonal elements follow $\rho(x, x', t) = \rho(x, x', 0)e^{-\Lambda t (x-x')^2}$, where the localization rate Λ is given by

$$\Lambda = k^2 \varphi \sigma_{\text{eff}} \qquad (552)$$

Ref. 732 where k is the wave number, φ the flux and σ_{eff} the cross section of the collisions, i.e. usually the size of the macroscopic object.

Ref. 733 One also finds the surprising result that a system hit by a particle of energy E_{hit} collapses the density matrix roughly down to the de Broglie (or thermal de Broglie) wavelength of the hitting particle. Both results together give the formula above.

Table 56 Common and less common baths with their main properties

Bath type	Temper-ature	Wave-length	Par-ticle flux	Hit time $t_{hit} = 1/\sigma\varphi$ for	
	T	λ_{eff}	φ	atom[a]	object[a]
matter baths					
solid, liquid	300 K	10 pm	10^{31} /m^{2}s	10^{-12} s	10^{-25} s
air	300 K	10 pm	10^{28} /m^{2}s	10^{-9} s	10^{-22} s
laboratory vacuum	50 mK	10 μm	10^{18} /m^{2}s	10 s	10^{-12} s
photon baths					
sunlight	5800 K	900 nm	10^{23} /m^{2}s	10^{-4} s	10^{-17} s
‘darkness’	300 K	20 μm	10^{21} /m^{2}s	10^{-2} s	10^{-15} s
cosmic microwaves	2.7 K	2 mm	10^{17} /m^{2}s	10^{2} s	10^{-11} s
terrestrial radio waves	.. K				
Casimir effect	.. K				
Unruh radiation of Earth	40 zK				
nuclear radiation baths					
radioactivity		10 pm		$10^{\cdot\cdot}$ s	$10^{\cdot\cdot}$ s
cosmic radiation	>1000 K	10 pm		$10^{\cdot\cdot}$ s	$10^{\cdot\cdot}$ s
solar neutrinos	≈ 10 MK	10 pm	10^{15} /m^{2}s	$10^{\cdot\cdot}$ s	$10^{\cdot\cdot}$ s
cosmic neutrinos	2.0 K	3 mm	10^{17} /m^{2}s	$10^{\cdot\cdot}$ s	$10^{\cdot\cdot}$ s
gravitational baths					
gravitational radiation	$5 \cdot 10^{31}$ K	10^{-35} m		>$10^{\cdot\cdot}$ s	>$10^{\cdot\cdot}$ s

a. The cross section σ in the case of matter and photon baths was assumed to be 10^{-19} m^2 for atoms; for the macroscopic object a size of 1 mm was used as example. For neutrino baths, ...

A second method to estimate the decoherence time is also common. Any interaction of a system with a bath is described by a relaxation time t_r. The term *relaxation* designates any process which leads to the return to the equilibrium state. The terms *damping* and *friction* are also used. In the present case, the relaxation time describes the return to equilibrium of the combination bath and system. Relaxation is an example of an irreversible evolution. A process is called *irreversible* if the reversed process, in which every component moves in opposite direction, is of very low probability.* For example, it is usual that a glass of wine poured into a bowl of water colours the whole water; it is very rarely observed that the wine and the water separate again, since the probability of all water and wine molecules to change directions together at the same time is rather low, a state of

* Beware of other definitions which try to make something deeper out of the concept of irreversibility, such as claims that ‘irreversible’ means that the reversed process is *not at all* possible. Many so-called ‘contradictions’ between the irreversibility of processes and the reversibility of evolution equations are due to this mistaken interpretation of the term ‘irreversible’.

affairs making the happiness of wine producers and the despair of wine consumers.

Now let us simplify the description of the bath. We approximate it by a single, unspecified, scalar field which interacts with the quantum system. Due to the continuity of space, such a field has an infinity of degrees of freedom. They are taken to model the many degrees of freedom of the bath. The field is assumed to be in an initial state where its degrees of freedom are excited in a way described by a temperature T. The interaction of the system with the bath, which is at the origin of the relaxation process, can be described by the repeated transfer of small amounts of energy E_{hit} until the relaxation process is completed.

The objects of interest in this discussion, like the mentioned cat, person or car, are described by a mass m. Their main characteristic is the maximum energy E_r which can be transferred from the system to the environment. This energy describes the interactions between system and environment. The superpositions of macroscopic states we are interested in are solutions of the Hamiltonian evolution of these systems.

The initial coherence of the superposition, so disturbingly in contrast with our everyday experience, disappears exponentially within a *decoherence time* t_d given by*

Ref. 734

$$t_d = t_r \frac{E_{\text{hit}}}{E_r} \frac{e^{E_{\text{hit}}/kT} - 1}{e^{E_{\text{hit}}/kT} + 1} \tag{555}$$

where k is the *Boltzmann constant* and like above, E_r is the maximum energy which can be transferred from the system to the environment. Note that one always has $t_d \leqslant t_r$. After the decoherence time t_d is elapsed, the system has evolved from the coherent to the incoherent superposition of states, or, in other words, the density matrix has lost its off-diagonal terms. One also says that the phase coherence of this system has been destroyed. Thus, after a time t_d, the system is found either in the state ψ_a or in the state ψ_b, respectively with the probability $|a|^2$ or $|b|^2$, and not any more in a coherent superposition which is so much in contradiction with our daily experience. Which final state is selected depends on the precise state of the bath, whose details were eliminated from the calculation by taking an *average* over the states of its microscopic constituents.

The important result is that for all macroscopic objects, the decoherence time t_d is extremely small. In order to see this more clearly, we can study a special simplified case. A macroscopic object of mass m, like the mentioned cat or car, is assumed to be at the same time in two locations separated by a distance l, i.e. in a superposition of the two corresponding states. We further assume that the superposition is due to the object mov-

* This result is derived as in the above case. A system interacting with a bath always has an evolution given by the general form

Ref. 735

$$\frac{d\rho}{dt} = -\frac{i}{\hbar}[H, \rho] - \frac{1}{2t_o} \sum_j [V_j \rho, V_j^\dagger] + [V_j, \rho V_j^\dagger] \, , \tag{553}$$

where ρ is the density matrix, H the Hamiltonian, V the interaction, and t_o the characteristic time of the interaction. Are you able to see why? Solving this equation, one finds for the elements far from the diagonal $\rho(t) = \rho_0 e^{-t/t_0}$. In other words, they disappear with a characteristic time t_o. In most situations one has a relation of the form

Challenge 1227 ny

$$t_0 = t_r \frac{E_{\text{hit}}}{E_r} = t_{\text{hit}} \tag{554}$$

or some variations of it, as in the example above.

ing as a quantum mechanical oscillator with frequency ω between the two locations; this is the simplest possible system that shows superpositions of an object located in two different positions. The energy of the object is then given by $E_r = m\omega^2 l^2$, and the smallest transfer energy $E_{\text{hit}} = \hbar\omega$ is the difference between the oscillator levels. In a macroscopic situation, this last energy is much smaller than kT, so that from the preceding expression
Ref. 736 we get

$$t_d = t_r \frac{E_{\text{hit}}^2}{2E_r\, kT} = t_r \frac{\hbar^2}{2mkTl^2} = t_r \frac{\lambda_T^2}{l^2} \qquad (556)$$

in which the frequency ω has disappeared. The quantity $\lambda_T = \hbar/\sqrt{2mkT}$ is called the *thermal de Broglie wavelength* of a particle.

It is straightforward to see that for practically all macroscopic objects the typical decoherence time t_d is extremely short. For example, setting $m = 1\,\text{g}$, $l = 1\,\text{mm}$ and $T = 300\,\text{K}$ we get $t_d/t_r = 1.3\cdot 10^{-39}$. Even if the interaction between the system and the environment would be so weak that the system would have as relaxation time the age of the universe, which is about $4\cdot 10^{17}\,\text{s}$, the time t_d would still be shorter than $5\cdot 10^{-22}\,\text{s}$, which is over a million times faster than the oscillation time of a beam of light (about 2 fs for green light). For Schrödinger's cat, the decoherence time would be even shorter. These times are so short that we cannot even hope to *prepare* the initial coherent superposition, let alone to observe its decay or to measure its lifetime.

For microscopic systems however, the situation is different. For example, for an electron in a solid cooled to liquid helium temperature we have $m = 9.1\cdot 10^{-31}\,\text{kg}$, and typically $l = 1\,\text{nm}$ and $T = 4\,\text{K}$; we then get $t_d \approx t_r$ and therefore the system can stay in a coherent superposition until it is relaxed, which confirms that for this case coherent effects can indeed be observed if the system is kept isolated. A typical example is the behaviour
Ref. 737 of electrons in superconducting materials. We will mention a few more below.

In 1996 the first actual measurement of decoherence times was published by the Paris
Ref. 738 team around Serge Haroche. It confirmed the relation between the decoherence time and the relaxation time, thus showing that the two processes have to be distinguished at microscopic scale. In the meantime, other experiments confirmed the decoherence
Ref. 739 process with its evolution equation, both for small and large values of t_d/t_r.

Conclusions on decoherence, life and death

In summary, both estimates of decoherence times tell us that for most macroscopic objects, in contrast to microscopic ones, both the preparation and the survival of superpositions of macroscopically different states is made practically impossible by the interaction with any bath found in their environment, even if the usual measure of this interaction, given by the friction of the motion of the system, is very small. Even if a macroscopic system is subject to an extremely low friction, leading to a very long relaxation time, its decoherence time is still vanishingly short.

Our everyday environment if full of baths. Therefore, *coherent superpositions of macroscopically distinct states never appear in everyday life.* In short, we cannot be dead and alive at the same time.

We also arrive at a second conclusion: *decoherence results from coupling to a bath in the environment.* Decoherence is a statistical or thermodynamic effect. We will return to this issue below.

What is a system? What is an object?

In classical physics, a system is a part of nature which can be isolated from its environment. However, quantum mechanics tells us that isolated systems do not exist, since interactions cannot be made vanishingly small. The results above allow us to define the concept of system with more accuracy. A *system* is any part of nature which interacts *incoherently* with its environment. In other words, an *object* is a part of nature interacting with its environment only through baths.

In particular, a system is called *microscopic* or *quantum mechanical* and can described by a wavefunction ψ whenever

- it is almost isolated, with $t_{\text{evol}} = \hbar/\Delta E < t_{\text{r}}$, and
- Ref. 740 it is in *incoherent* interaction with its environment.

In short, a microscopic system interacts incoherently and weakly with its environment. In contrast, a bath is never isolated in the sense just given, because its evolution time is always much larger than its relaxation time. Since all macroscopic bodies are in contact with baths – or even contain one – they cannot be described by a wavefunction. In particular, it is impossible to describe any measuring apparatus with the help of a wavefunction.

We thus conclude that a *macroscopic system* is a system with a decoherence time much shorter than any other evolution time of its constituents. Obviously, macroscopic systems also interact incoherently with their environment. Thus cats, cars and television news speakers are all macroscopic systems.

A third possibility is left over by the two definitions: what happens in the situation in which the interactions with the environment are *coherent*? We will encounter some examples shortly. Following this definition, such situations are *not* systems and cannot be described by a wavefunction. For example, it can happen that a particle forms neither a macroscopic nor a microscopic system! In these situations, when the interaction is coherent, one speaks of *entanglement*; such a particle or set of particles is said to be entangled with its environment.

Entangled, coherently interacting systems are separable, but not divisible. In quantum theory, nature is not found to be made of isolated entities, but is still made of *separable* entities. The criterion of separability is the incoherence of interaction. Coherent superpositions imply the surprising consequence that there are systems which, even though they look divisible, are not. Entanglement poses a limit to divisibility. All surprising properties of quantum mechanics, such as Schrödinger's cat, are consequences of the classical prejudice that a system made of two or more parts must necessarily be divisible into two subsystems. But coherent superpositions, or entangled systems, do not allow division. Whenever we try to divide indivisible systems, we get strange or incorrect conclusions, such as apparent faster-than-light propagation, or, as one says today, non-local behaviour. Let us have a look at a few typical examples.

Is quantum theory non-local? – A bit about Einstein-Podolsky-Rosen

[Mr. Duffy] lived a little distance away from his body ...
James Joyce, *A Painful Case*

Page 457 After we explored non-locality in general relativity, we now study it in quantum mechanics. We first look at the wavefunction collapse for an electron hitting a screen after passing a slit. Following the description just deduced, the process proceeds schematically as depicted in Figure 312. A movie of the same process can be seen in the lower left corners on the pages following page 655. The situation is surprising: due to the short decoherence time, in a wavefunction collapse the maximum of the function changes position at extremely high speed. In fact, the maximum moves faster than light. But is it a problem?

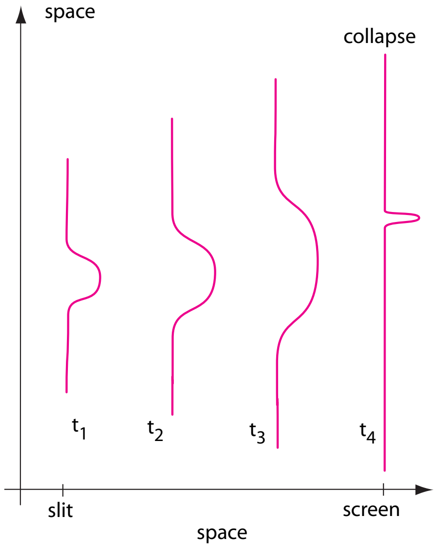


Figure 312 Quantum mechanical motion: an electron wave function (actually its module squared) from the moment it passes a slit until it hits a screen

A situation is called *acausal* or *non-local* if energy is transported faster than light. Using Challenge 1228 n Figure 312 you can determine the energy velocity Page 533 involved, using the results on signal propagation. The result is a value smaller than c. A wavefunction maximum moving faster than light does *not* imply energy moving faster than light.

In other words, quantum theory has speeds Ref. 741 greater than light, but no energy speeds greater than light. In classical electrodynamics, the same happens with the scalar and the vector potentials if the Coulomb gauge is used. We have also encountered speeds faster than that of light in the motion of shadows and Page 278 in many other observations. Any physicist now has two choices: he can be straight, and say that there is no non-locality in nature; or he can be less straight, and claim there is. In the latter case, he has to claim that even classical physics is non-local. However, this never happens. On the other hand, there is a danger in this more provoking usage: a small percentage of those who say that the world is non-local after a while start to believe that there really are faster-than-light effects in nature. These people become prisoners of their muddled thinking; on the other hands, muddled thinking helps to get more easily into newspapers. In short, even though the definition of non-locality is not unanimous, here we stick to the stricter one, and define non-locality as energy transport faster than light.

An often cited Gedanken experiment that shows the pitfalls of non-locality was proposed by Bohm* in the discussion around the so-called Einstein–Podolsky–Rosen Ref. 742, Ref. 743 paradox. In the famous EPR paper the three authors try to find a contradiction between quantum mechanics and common sense. Bohm translated their rather confused paper into a clear Gedanken experiment. When two particles in a spin 0 state move apart,

* David Joseph Bohm (1917–1992) American–British physicist, codiscovered the Aharonov–Bohm effect; he spent a large part of his life investigating the connections between quantum physics and philosophy.

measuring one particle's spin orientation implies an *immediate* collapse also of the other particle's spin, namely in the exactly opposite direction. This happens instantaneously over the whole separation distance; no speed limit is obeyed. In other words, entanglement seems to lead to faster-than-light communication.

Again, in Bohm's experiment, no energy is transported faster than light. No non-locality is present, despite numerous claims of the contrary by certain authors. The two entangled electrons belong to one system: assuming that they are separate only because the wavefunction has two distant maxima is a conceptual mistake. In fact, no signal can be transmitted with this method; the decoherence is a case of prediction which looks like a signal without being one. We already discussed such cases in the section on Page 537 electrodynamics.

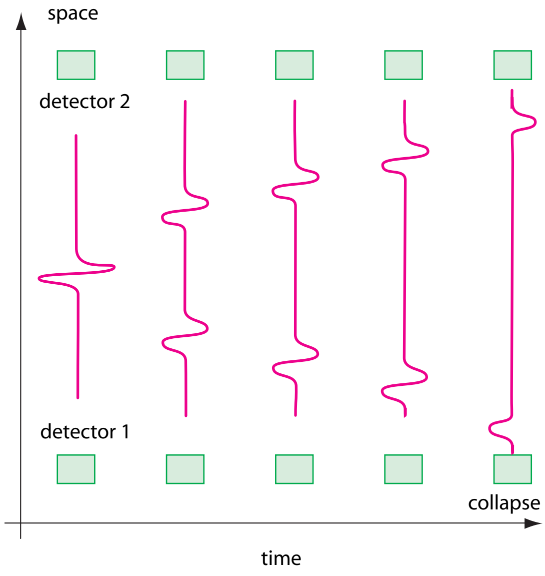


Figure 313 Bohm's Gedanken experiment

Bohm's experiment has actually been performed. The first and most famous was realized in 1982 by Alain Aspect and used photons instead of Ref. 744 electrons. Like all latter tests, it has fully confirmed quantum mechanics.

In fact, experiments such as the one by Aspect confirm that it is not possible to treat either of the two particles as a system by itself and to ascribe either of them any property, such as a spin orientation, only by itself. The Heisenberg picture would express this even more clearly.

The mentioned two examples of apparent non-locality can be dismissed with the remark that since obviously no energy flux faster than light is involved, no problems with causality appear. Therefore the following example is more interesting. Take two identical atoms, one in an excited state, one in the ground state, and call l the distance that separates them. Common sense tells that if the first atom returns to its ground state emitting a photon, the second atom can be excited only after a time $t = l/c$ has been elapsed, i.e. after the photon has travelled to the second atom.

Surprisingly, this conclusion is wrong. The atom in its ground state has a non-zero probability to be excited at the same moment in which the first is de-excited. This has been Ref. 745 shown most simply by Hegerfeldt. The result has even been confirmed experimentally.

More careful studies show that the result depends on the type of superposition of the two atoms at the beginning: coherent or incoherent. For incoherent superpositions, the intuitive result is correct; the surprising result appears only for coherent superpositions. The discussion again shows that no real non-locality of energy is involved.

Curiosities

Coherent superposition, or entanglement, is such a surprising phenomenon that many aspects have been and still are being explored.

▪ In a few rare cases, the superposition of different macroscopic states can actually be observed by lowering the temperature to sufficiently small values and by carefully choosing suitably small masses or distances. Two well-known examples of coherent superpositions are those observed in gravitational wave detectors and in Josephson junctions. In Ref. 736 the first case, one observes a mass as heavy as 1000 kg in a superposition of states located at different points in space: the distance between them is of the order of 10^{-17} m. In the second case, in superconducting rings, superpositions of a state in which a macroscopic current of the order of 1 pA flows in clockwise direction with one where it flows in Ref. 753 counter-clockwise direction have been produced.

▪ Superpositions of magnetization in up and down direction at the same time have Ref. 747 also be observed for several materials.

▪ Since the 1990s, the sport of finding and playing with new systems in coherent macro- Ref. 749 scopic superpositions has taken off across the world. The challenges lie in the clean experiments necessary. Experiments with single atoms in superpositions of states are among Ref. 750 the most popular ones.

Ref. 751 ▪ In 1997, coherent atom waves were extracted from a cloud of sodium atoms.

▪ Macroscopic objects usually are in incoherent states. This is the same situation as for light. The world is full of 'macroscopic', i.e. incoherent light: daylight, and all light from lamps, from fire and from glow-worms is incoherent. Only very special and carefully constructed sources, such as lasers or small point sources, emit coherent light. Only these allow to study interference effects. In fact, the terms 'coherent' and 'incoherent' originated in optics, since for light the difference between the two, namely the capacity to interfere, had been observed centuries before the case of matter.

Coherence and incoherence of light and of matter manifest themselves differently, since matter can stay at rest but light cannot and because light is made of bosons, but Page 735 matter is made of fermions. Coherence can be observed easily in systems composed of bosons, such as light, sound in solids, or electron pairs in superconductors. Coherence is less easily observed in systems of fermions, such as systems of atoms with their electron clouds. However, in both cases a decoherence time can be defined. In both cases coherence in many particle systems is best observed if all particles are in the same state (superconductivity, laser light) and in both cases the transition from coherent to incoherent is due to the interaction with a bath. A beam is thus incoherent if its particles arrive randomly in time and in frequency. In everyday life, the rarity of observation of coherent matter superpositions has the same origin as the rarity of observation of coherent light.

▪ We will discuss the relation between the environment and the *decay* of unstable systems later on. The phenomenon is completely described by the concepts given here.

Challenge 1229 ny ▪ Can you find a method to measure the *degree* of entanglement? Can you do so for a system made of many particles?

▪ The study of entanglement leads to a simple conclusion: *teleportation contradicts cor-* Challenge 1230 ny *relation*. Can you confirm the statement?

▪ Some people say that quantum theory could be used for quantum computing, by us- Ref. 755 ing coherent superpositions of wavefunctions. Can you give a general reason that makes

Challenge 1231 n this aim very difficult, even without knowing how such a quantum computer might work?

What is all the fuss about measurements in quantum theory?

Measurements in quantum mechanics are disturbing. They lead to statements in which *probabilities* appear. That is puzzling. For example, we speak about the probability of finding an electron at a certain distance from the nucleus of an atom. Statements like this belong to the general type 'when the observable A is measured, the probability to find the outcome a is p.' In the following we will show that the probabilities in such statements are inevitable for any measurement, because, as we will show, any measurement and any observation is a special case of decoherence or disentanglement process. (Historically however, the process of measurement was studied before the more general process of decoherence. That explains in part why the topic is so confused in many peoples' minds.)

Page 591 What is a measurement? As already mentioned in the intermezzo a measurement is any interaction which produces a record or a memory. (In everyday life, any effect is a record; but this is not true in general. Can you give some examples and counter-examples?) Challenge 1232 ny Measurements can be performed by machines; when they are performed by people, they are called observations. In quantum theory, the action of measurement is not as straightforward as in classical physics. This is seen most strikingly when a quantum system, such as a single electron, is first made to pass a diffraction slit, or better – in order to make its wave aspect become apparent – a double slit and then is made to hit a photographic plate, in order to make also its particle aspect appear. Experiment shows that the blackened dot, the spot where the electron has hit the screen, cannot be determined in advance. (The same is true for photons or any other particle.) However, for large numbers of electrons, the spatial distribution of the black dots, the so-called *diffraction pattern*, can be calculated in advance with high precision.

The outcome of experiments on microscopic systems thus forces us to use probabilities for the description of microsystems. We find that the probability distribution $p(\mathbf{x})$ of the spots on the photographic plate can be calculated from the wavefunction ψ of the electron at the screen surface and is given by $p(\mathbf{x}) = |\psi^\dagger(\mathbf{x})\psi(\mathbf{x})|^2$. This is in fact a special case of the general *first property of quantum measurements*: the measurement of an observable A for a system in a state ψ gives as result one of the eigenvalues a_n, and the probability P_n to get the result a_n is given by

$$P_n = |\varphi_n^\dagger \psi|^2 , \qquad (557)$$

Page 1110 where φ_n is the eigenfunction of the operator A corresponding to the eigenvalue a_n.

Experiments also show a *second property of quantum measurements:* after the measurement, the observed quantum system is in the state φ_n corresponding to the measured eigenvalue a_n. One also says that during the measurement, the wavefunction has *collapsed* Ref. 752 from ψ to φ_n. By the way, both properties can also be generalized to the more general cases with degenerate and continuous eigenvalues.

At first sight, the sort of probabilities encountered in quantum theory are different from the probabilities we encounter in everyday life. Roulette, dice, pachinko machines, the direction in which a pencil on its tip falls, have been measured experimentally to be random (assuming no cheating by the designer or operators) to a high degree of accuracy.

These systems do not puzzle us. We unconsciously assume that the random outcome is due to the small, but uncontrollable variations of the starting conditions every time the experiment is repeated.*

But microscopic systems seem to be different. The two measurement properties just mentioned express what physicists observe in every experiment, even if the initial conditions are taken to be *exactly* the same every time. But why then is the position for a single electron, or most other observables of quantum systems, not predictable? In other words, what happens during the collapse of the wavefunction? How long does it take? In the beginning of quantum theory, there was the perception that the observed unpredictability is due to the lack of information about the state of the particle. This lead many to search for so-called 'hidden variables'. All these attempts were doomed to fail, however. It took some time for the scientific community to realize that the unpredictability is *not* due to the lack of information about the state of the particle, which is indeed described *completely* by the state vector ψ.

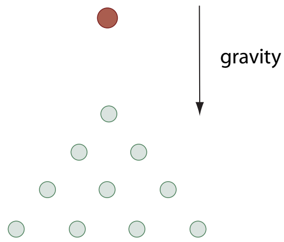


Figure 314 A system showing probabilistic behaviour

In order to uncover the origin of probabilities, let us recall the nature of a measurement, or better, of a general observation. *Any observation is the production of a record.* The record can be a visual or auditive memory in our brain, or a written record on paper, or a tape recording, or any such type of object. As explained in the intermezzo, an Page 590 object is a record if it cannot have arisen or disappeared by chance. To avoid the influence of chance, all records have to be protected as much as possible from the outer world; e.g. one typically puts archives in earthquake safe buildings with fire protection, keeps documents in a safe, avoids brain injury as much as possible, etc.

On top of this, records have to be protected from their internal fluctuations. These internal fluctuations are due to the many components any recording device is made of. But if the fluctuations were too large, they would make it impossible to distinguish between the possible contents of a memory. Now, fluctuations decrease with increasing size of a system, typically with the square root of the size. For example, if a hand writing is too small, it is difficult to read if the paper gets brittle; if the magnetic tracks on tapes are too small, they demagnetize and loose the stored information. In other words, a record is rendered stable against internal fluctuations by making it of sufficient size. Every record thus consists of many components and shows small fluctuations.

Therefore, every system with memory, i.e. every system capable of producing a record, contains a *bath*. In summary, the statement that any observation is the production of a record can be expressed more precisely as: *Any observation of a system is the result of an interaction between that system and a bath in the recording apparatus.**

But we can say more. Obviously, any observation measuring a physical quantity uses an interaction *depending* on that same quantity. With these seemingly trivial remarks, we can describe in more detail the process of observation, or, as it is usually called in the quantum theory, the measurement process.

* To get a feeling for the limitations of these unconscious assumptions, you may want to read the already mentioned story of those physicists who built a machine that could predict the outcome of a roulette ball from the initial velocity imparted by the croupier. Ref. 79

* Since baths imply friction, we can also say: memory needs friction.

Any measurement apparatus, or *detector*, is characterized by two main aspects: the interaction it has with the microscopic system, and the bath it contains to produce the record. Any description of the measurement process thus is the description of the evolution of the microscopic system *and* the detector; therefore one needs the Hamiltonian for the particle, the interaction Hamiltonian, and the bath properties, such as the relaxation time. The interaction specifies what is measured and the bath realizes the memory.

Ref. 748

We know that only classical thermodynamic systems can be irreversible; quantum systems are not. We therefore conclude: a measurement system *must* be described classically: otherwise it has no memory and is not a measurement system: it produces no record! Nevertheless, let us see what happens if one describes the measurement system quantum mechanically. Let us call A the observable which is measured in the experiment and its eigenfunctions φ_n. We describe the quantum mechanical system under observation – often a particle – by a state ψ. This state can always be written as $\psi = \psi_p \psi_{\text{other}} = \sum_n c_n \varphi_n \psi_{\text{other}}$, where ψ_{other} represents the other degrees of freedom of the particle, i.e. those not described – *spanned*, in mathematical language – by the operator A corresponding to the observable we want to measure. The numbers $c_n = |\varphi_n^\dagger \psi_p|$ give the expansion of the state ψ_p, which is taken to be normalized, in terms of the basis φ_n. For example, in a typical position measurement, the functions φ_n would be the position eigenfunctions and ψ_{other} would contain the information about the momentum, the spin and all other properties of the particle.

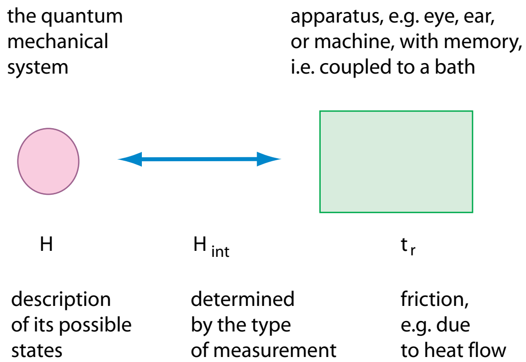


Figure 315 The concepts used in the description of measurements

How does the system–detector interaction look like? Let us call the state of the apparatus before the measurement χ_{start}; the measurement apparatus itself, by definition, is a device which, when it is hit by a particle in the state $\varphi_n \psi_{\text{other}}$, changes from the state χ_{start} to the state χ_n. One then says that the apparatus has *measured* the eigenvalue a_n corresponding to the eigenfunction φ_n of the operator A. The index n is thus the record of the measurement; it is called the *pointer* index or variable. This index tells us in which state the microscopic system was before the interaction. The important point, taken from our previous discussion, is that the states χ_n, being records, are macroscopically distinct, precisely in the sense of the previous section. Otherwise they would not be records, and the interaction with the detector would not be a measurement.

Of course, during measurement, the apparatus sensitive to φ_n changes the part ψ_{other} of the particle state to some other situation $\psi_{\text{other},n}$, which depends on the measurement and on the apparatus; we do not need to specify it in the following discussion.* Let us have an intermediate check of our reasoning. Do apparatuses as described here exist? Yes, they do. For example, any photographic plate is a detector for the position of ionizing particles. A plate, and in general any apparatus measuring position, does this by changing its momentum in a way depending on the measured position: the electron on a photographic plate is stopped. In this case, χ_{start} is a white plate, φ_n would be a particle localized at spot n, χ_n is the function describing a plate blackened at spot n and $\psi_{\text{other,n}}$ describes the momentum and spin of the particle after it has hit the photographic plate at the spot n.

Now we are ready to look at the measurement process itself. For the moment, let us disregard the bath in the detector. In the time before the interaction between the particle and the detector, the combined system was in the initial state ψ_i given simply by

$$\psi_i = \psi_p \chi_{\text{start}} = \sum_n c_n \varphi_n \psi_{\text{other}} \chi_{\text{start}} \ . \tag{560}$$

After the interaction, using the just mentioned characteristics of the apparatus, the combined state ψ_a is

$$\psi_a = \sum_n c_n \varphi_n \psi_{\text{other},n} \chi_n \ . \tag{561}$$

This evolution from ψ_i to ψ_a follows from the evolution equation applied to the particle detector combination. Now the state ψ_a is a superposition of macroscopically distinct states, as it is a superposition of distinct macroscopic states of the detector. In our example ψ_a could correspond to a superposition of a state where a spot on the left upper corner is blackened on an otherwise white plate with one where a spot on the right lower corner of the otherwise white plate is blackened. Such a situation is never observed. Let us see why. The density matrix ρ_a of this situation, given by

$$\rho_a = \psi_a \otimes \psi_a^\dagger = \sum_{n,m} c_n c_m^* \left(\varphi_n \psi_{\text{other},n} \chi_n\right) \otimes \left(\varphi_m \psi_{\text{other},m} \chi_m\right)^\dagger \ , \tag{562}$$

contains non-diagonal terms, i.e. terms for $n \neq m$, whose numerical coefficients are different from zero. Now let's take the bath back in.

From the previous section we know the effect of a bath on such a macroscopic superposition. We found that a density matrix such as ρ_a decoheres extremely rapidly. We assume here that the decoherence time is negligibly small, in practice thus instantaneous,*

* How does the interaction look like mathematically? From the description we just gave, we specified the final state for every initial state. Since the two density matrices are related by

$$\rho_{\text{f}} = T \rho_{\text{i}} T^\dagger \tag{558}$$

Challenge 1233 ny we can deduce the Hamiltonian from the matrix T. Are you able to see how?

By the way, one can say in general that an apparatus measuring an observable A has a system interaction Hamiltonian depending on the pointer variable A, and for which one has

$$[H + H_{\text{int}}, A] = 0 \ . \tag{559}$$

* Note however, that an *exactly* vanishing decoherence time, which would mean a *strictly* infinite number

so that the off-diagonal terms vanish, and only the final, diagonal density matrix ρ_f, given by

$$\rho_\mathrm{f} = \sum_n |c_n|^2 (\varphi_n \psi_{\mathrm{other},n} \chi_n) \otimes (\varphi_n \psi_{\mathrm{other},n} \chi_n)^\dagger \tag{563}$$

has experimental relevance. As explained above, such a density matrix describes a mixed state and the numbers $P_n = |c_n|^2 = |\varphi_n^\dagger \psi_p|^2$ give the probability of measuring the value a_n and of finding the particle in the state $\varphi_n \psi_{\mathrm{other,n}}$ as well as the detector in the state χ_n. But this is precisely what the two properties of quantum measurements state.

We therefore find that describing a measurement as an evolution of a quantum system interacting with a macroscopic detector, itself containing a bath, we can *deduce* the two properties of quantum measurements, and thus the collapse of the wave function, from the quantum mechanical evolution equation. The decoherence time t_d of the previous section becomes the time of collapse in the case of a measurement:

$$t_\mathrm{collapse} = t_\mathrm{d} < t_\mathrm{r} \tag{564}$$

We thus have a formula for the time the wavefunction takes to collapse. The first experimental measurements of the time of collapse are appearing and confirm these results. Ref. 754

Hidden variables

Obviously a large number of people are not satisfied with the arguments just presented. They long for more mystery in quantum theory. The most famous approach is the idea that the probabilities are due to some hidden aspect of nature which is still unknown to humans. But the beautiful thing about quantum mechanics is that it allows both conceptual and experimental tests on whether such *hidden variables* exist without the need of knowing them.

- Clearly, hidden variables controlling the evolution of microscopic system would contradict the result that action values below $\hbar/2$ cannot be detected. This minimum observable action is the reason for the random behaviour of microscopic systems.
- Historically, the first argument against hidden variables was given by John von Neumann.*

– CS – to be written – CS –

- An additional no-go theorem for hidden variables was published by Kochen and Specker in 1967, (and independently by Bell in 1969). It states that non-contextual hidden variables are impossible, if the Hilbert space has a dimension equal or larger than Ref. 756

of degrees of freedom of the environment, is in contradiction with the evolution equation, and in particular with unitarity, locality and causality. It is essential in the whole argument not to confuse the logical consequences of a extremely small decoherence time with those of an exactly vanishing decoherence time.

* János von Neumann (b. 1903 Budapest, d. 1957 Washington DC) Hungarian mathematician. One of the greatest and clearest minds of the twentieth century, he settled already many questions, especially in applied mathematics and quantum theory, that others still struggle with today. He worked on the atomic and the hydrogen bomb, on ballistic missiles, and on general defence problems. In another famous project, he build the first US-American computer, building on his extension of the ideas of Konrad Zuse.

three. The theorem is about non-contextual variables, i.e. about hidden variables *inside* the quantum mechanical system. The Kochen–Specker theorem thus states that there is no non-contextual hidden variables model, because mathematics forbids it. This result essentially eliminates all possibilities, because usual quantum mechanical systems have dimensions much larger than three.

But also common sense eliminates hidden variables, without any recourse to mathematics, with an argument often overlooked. If a quantum mechanical system had internal hidden variables, the measurement apparatus would have zillions of them.* And that would mean that it could not work as a measurement system.

Of course, one cannot avoid noting that about *contextual* hidden variables, i.e. variables in the environment, there are no restricting theorems; indeed, their necessity was shown earlier in this section.

▪ Obviously, despite these results, people have also looked for experimental tests on hidden variables. Most tests are based on the famed *Bell's equation*, a beautifully simple relation published by John Bell** in the 1960s.

The starting idea is to distinguish quantum theory and locally realistic theories using hidden variables by measuring the polarizations of two correlated photons. Quantum theory says that the polarization of the photons is fixed only at the time it is measured, whereas local realistic theories say that it is fixed already in advance. The correct description can be found by experiment.

Imagine the polarization is measured at two distant points A and B, each observer can measure 1 or −1 in each of his favourite direction. Let each observer choose two directions, 1 and 2, and call their results a_1, a_2, b_1 and b_2. Since the measurement results all are either 1 or −1, the value of the specific expression $(a_1 + a_2)b_1 + (a_2 - a_1)b_2$ has always the value ±2.

Ref. 757 Imagine you repeat the experiment many times, assuming that the hidden variables appear statistically. You then can deduce (a special case of) Bell's inequality for two hidden Challenge 1234 e variables

$$|(a_1 b_1) + (a_2 b_1) + (a_2 b_2) - (a_1 b_2)| \leqslant 2 \tag{565}$$

where the expressions in brackets are the averages of the measurement products over a large number of samples. This result holds independently of the directions of the involved polarizers.

On the other hand, if the polarizers 1 and 2 at position A and the corresponding ones at position B are chosen with angles of $\pi/4$, quantum theory predicts that the result is

$$|(a_1 b_1) + (a_2 b_1) + (a_2 b_2) - (a_1 b_2)| = 2\sqrt{2} > 2 \tag{566}$$

which is in complete contradiction with the hidden variable result.

So far, all experimental checks of Bell's equation have confirmed standard quantum mechanics. No evidence for hidden variables has been found. This is not really surprising, since the search for such variables is based on a misunderstanding of quantum mechanics

* Which leads to the definition: one zillion is 10^{23}.

** John Stewart Bell (1928–1990), theoretical physicist who worked mainly on the foundations of quantum theory.

or on personal desires on how the world should be, instead of relying on experimental evidence.

Another measurable contradiction between quantum theory and locally realistic theories has been predicted by Greenberger, Horn and Zeilinger. Experiments trying to check the result are being planned. No deviation from quantum theory is expected.

Conclusions on probabilities and determinism

> Geometric demonstramus quia facimus; si physics demonstrare possemus, faceremus.
>
> Giambattista Vico*

From the arguments presented here we draw a number of conclusions which we need for the rest of our mountain ascent. Note that these conclusions are not yet shared by all physicists! The whole topic is still touchy.

▪ Probabilities do not appear in measurements because the state of the quantum system is unknown or fuzzy, but because the detailed state of the bath in the environment is unknown. *Quantum mechanical probabilities are of statistical origin and are due to baths in the environment (or the measurement apparatus).* The probabilities are due to the large number of degrees of freedom contained in any bath. These large numbers make the outcome of experiments unpredictable. If the state of the bath were known, the outcome of an experiment could be predicted. The probabilities of quantum theory are 'thermodynamic' in origin.

In other words, there are *no* fundamental probabilities in nature. All probabilities in nature are due to decoherence; in particular, all probabilities are due to the statistics of the many particles – some of which may be virtual – that are part of the baths in the environment. Modifying well-known words by Albert Einstein, 'nature really does not play dice.' We therefore called ψ the *wave function* instead of 'probability amplitude', as is often done. *State function* would be an even better name.

▪ Any observation in everyday life is a special case of decoherence. What is usually called the 'collapse of the wavefunction' is a decoherence process due to the interaction with the baths present in the environment or in the measuring apparatus. Because humans are warm-blooded and have memory, humans themselves are thus measurement apparatuses. The fact that our body temperature is 37°C is thus the reason that we see only a single world, and no superpositions. (Actually, there are more reasons; can you name a few?) Challenge 1235 n

▪ A measurement is complete when the microscopic system has interacted with the bath in the measuring apparatus. Quantum theory as a description of nature does not require detectors; the evolution equation describes all examples of motion. However, *measurements* do require the existence of detectors. Detectors, being machines that record observations, have to include a bath, i.e. have to be classical, macroscopic objects. In this

* 'We are able to demonstrate geometrical matters because we make them; if we could prove physical matters we would be able to make them.' Giovanni Battista Vico (b. 1668 Napoli, d. 1744 Napoli) important Italian philosopher and thinker. In this famous statement he points out a fundamental distinction between mathematics and physics.

context one speaks also of a *classical apparatus*. This necessity of the measurement apparatus to be classical had been already stressed in the very early stages of quantum theory.

▪ All measurements, being decoherence processes that involve interactions with baths, are irreversible processes and increase entropy.

▪ A measurement is a special case of quantum mechanical evolution, namely the evolution for the combination of a quantum system, a macroscopic detector and the environment. Since the evolution equation is relativistically invariant, no causality problems appear in measurements; neither do locality problems and logical problems appear.

▪ Since both the evolution equation and the measurement process does not involve quantities other than space-time, Hamiltonians, baths and wave-functions, no other quantity plays a role in measurement. In particular, no human observer nor any consciousness are involved or necessary. Every measurement is complete when the microscopic system has interacted with the bath in the apparatus. The decoherence inherent in every measurement takes place even if nobody is looking. This trivial consequence is in agreement with the observations of everyday life, for example with the fact that the Moon is orbiting the Earth even if nobody looks at it.* Similarly, a tree falling in the middle of a forest makes noise even if nobody listens. Decoherence is independent of human observation, of the human mind and of human existence.

▪ In every measurement the quantum system interacts with the detector. Since there is a minimum value for the magnitude of action, every *observation influences the observed*. Therefore every measurement *disturbs* the quantum system. Any precise description of observations must also include the description of this disturbance. In the present section the disturbance was modelled by the change of the state of the system from ψ_{other} to $\psi_{\text{other,n}}$. Without such a change of state, without a disturbance of the quantum system, a measurement is impossible.

▪ Since the complete measurement is described by quantum mechanics, unitarity is and remains the basic property of evolution. There are no non-unitary processes in quantum mechanics.

▪ The description of the collapse of the wavefunction as a decoherence process is an explanation exactly in the sense in which the term 'explanation' was defined in the Page 639 intermezzo; it describes the relation between an observation and all the other aspects of reality, in this case the bath in the detector or the environment. The collapse of the wavefunction has been both calculated and *explained*. The collapse is not a question of 'interpretation', i.e. of opinion, as unfortunately often is suggested.**

▪ It is not useful to speculate whether the evolution for a *single* quantum measurement could be determined if the state of the environment around the system were known. Measurements need baths. But baths cannot be described by wavefunctions.*** Quantum mechanics is deterministic. Baths are probabilistic.

* The opposite view is sometimes falsely attributed to Niels Bohr. The Moon is obviously in contact with many radiation baths. Can you list a few? Challenge 1236 n

** This implies that the so-called 'many worlds' interpretation is wishful thinking. The conclusion is confirmed when studying the details of this religious approach. It is a belief system, not based on facts. Ref. 758

*** This very strong type of determinism will be very much challenged in the last part of this text, in which it will be shown that time is not a fundamental concept, and therefore that the debate around determinism looses most of its interest.

▪ In summary, there is *no* irrationality in quantum theory. Whoever uses quantum theory as argument for superstitions, irrational behaviour, new age beliefs or ideologies is guilty of disinformation. The statement by Gell-Mann at the beginning of this chapter is thus such an example. Another is the following well-known but incorrect statement by Richard Feynman:

Ref. 759 ...nobody understands quantum mechanics.

Nobel prizes in physics obviously do not prevent infection by ideology. On the other hand, the process of decoherence allows a clear look on various interesting issues.

What is the difference between space and time?

Space and time differ. Objects are localized in space but not in time. Why is this the case? Most bath–system interactions are mediated by a potential. All potentials are by definition position dependent. Therefore, every potential, being a function of the position $\mathbf{x}$, commutes with the position observable (and thus with the interaction Hamiltonian). The decoherence induced by baths – except if special care is taken – thus first of all destroys the non-diagonal elements for every superposition of states centred at different locations. In short, *objects are localized because they interact with baths via potentials.*

For the same reason, objects also have only one spatial orientation at a time. If the system–bath interaction is spin-dependent, the bath leads to 'localization' in the spin variable. This happens for all microscopic systems interacting with magnets. As a result, macroscopic superpositions of magnetization are almost never observed. Since electrons, protons and neutrons have a magnetic moment and a spin, this conclusion can even be extended: everyday objects are never seen in superpositions of different rotation states because their interactions with baths are spin-dependent.

As a counterexample, most systems are not localized in time, but on the contrary exist for very long times, because practically all system–bath interactions do *not* commute with time. In fact, this is the way a bath is defined to begin with. In short, *objects are permanent because they interact with baths.*

Are you able to find an interaction which is momentum-dependent? What is the consequence for macroscopic systems? Challenge 1237 n

In other words, in contrast to general relativity, quantum theory produces a distinction between space and time. In fact, we can *define* position as the observable that commutes with interaction Hamiltonians. This distinction between space and time is due to the properties of matter and its interactions. We could not have deduced this distinction in general relativity.

Are we good observers?

Are humans classical apparatuses? Yes, they are. Even though several prominent physicists claim that free will and probabilities are related, a detailed investigation shows that this in not the case. Our senses are classical machines in the sense described above: they record observations by interaction with a bath. Our brain is also a classical apparatus, but the fact is secondary; our sensors are the key. Ref. 760

Any observing entity needs a bath and a memory to record its observations. That means that observers have to be made of matter; an observer cannot be made of radiation. Our description of nature is thus severely biased: we describe it from the standpoint of matter. That is a bit like describing the stars by putting the Earth at the centre of the universe. Can we eliminate this basic anthropomorphism? We will find out as we continue.

What connects information theory, cryptology and quantum theory?

Physics means talking about observations of nature. Like any observation, also measurements produce information. It is thus possible to translate much (but not all) of quantum theory into the language of information theory. In particular, the existence of a minimal change in nature implies that the information about a physical system can never be complete, that information transport has its limits and that information can never be fully trusted. The details of these studies form a fascinating way to look at the microscopic world. The studies become even more interesting when the statements are translated into
Ref. 761 the language of cryptology. Cryptology is the science of transmitting hidden messages that only the intended receiver can decrypt. In our modern times of constant surveillance, cryptology is an important tool to protect personal freedom.*

The quantum of action implies that messages can be sent in an (almost) safe way. Listening to a message is a measurement process. Since there is a smallest action, one can detect whether somebody has tried to listen to a sent message. A man in the middle attack – somebody who pretends to be the receiver and then sends a copy of the message to the real, intended receiver – can be avoided by using entangled systems as signals to transmit the information. Quantum cryptologists therefore usually use communication systems based on entangled photons.

The major issue of quantum cryptology is the key distribution problem. All secure communication is based on a secret key that is used to decrypt the message. Even if the communication channel is of the highest security – like entangled photons – one still has to find a way to send the communication partner the secret key necessary for the decryption of the messages. Finding such methods is the main aspect of quantum cryptology, a large research field. However, close investigation shows that all key exchange methods are limited in their security. In short, due to the quantum of action, nature provides limits on the possibility of sending encrypted messages. The statement of these limits is (almost) equivalent to the statement that change in nature is limited by the quantum of action.

The quantum of action provides a limit to secure information exchange. This connection also allows to brush aside several incorrect statements often found in the media.
Page 968 Stating that 'the universe is information' or that 'the universe is a computer' is as devoid of reason as saying that the universe is an observation, a measurement apparatus, a clockwork or a chewing-gum dispenser. Any expert of motion should beware of these and similarly fishy statements; people who use them either deceive themselves or try to deceive others.

* Cryptology consists of the field of *cryptography*, the art of coding messages, and the field of *cryptoanalysis*, the art of deciphering encrypted messages. For a good introduction to cryptology, see the text by ALBRECHT BEUTELSPACHER, JÖRG SCHWENK & KLAUS-DIETER WOLFENSTÄTTER, *Moderne Verfahren der Kryptographie*, Vieweg 1995.

Does the universe have a wave function? And initial conditions?

The wavefunction of the universe is is frequently invoked in discussions about quantum theory. Many deduce conclusions from it, for example on the irreversibility of time, on the importance of initial conditions, on changes required to quantum theory and much more. Are these arguments correct?

Page 409 The first thing to clarify is the meaning of 'universe'. As explained above, the term can have two meanings: either the collection of all matter and radiation, or this collection *plus* all of space-time. Then we recall the meaning of 'wavefunction': it describes the *state* of a system. The state distinguishes two otherwise identical systems; for example, position and velocity distinguish two otherwise identical ivory balls on a billiard table. Alternatively and equivalently, the state describes changes in time.

Does the universe have a state? If we take the wider meaning of universe, obviously Page 38 it does not. Talking about the state of the universe is a contradiction: by definition, the concept of state, defined as the non-permanent aspects of an object, is applicable only to *parts* of the universe.

We then can take the narrower sense of 'universe' – the sum of all matter and radiation only – and ask the question again. To determine its state, we need a possibility to measure it: we need an environment. But the environment of matter and radiation is space-time only; initial conditions cannot be determined since we need measurements to do this, and thus an apparatus. An apparatus is material system with a bath attached to it; there is no such system outside the universe.

In short, quantum theory does not allow for measurements of the universe; therefore the universe has no state. Beware of anybody who claims to know something about the wavefunction of the universe. Just ask him: If you know the wavefunction of the universe, why aren't you rich?

Despite this conclusion, several famous physicists have proposed evolution equations Ref. 762 for the wavefunction of the universe. (The best-known is the Wheeler–DeWitt equation.) It seems a silly point, but the predictions of these equations have not been compared to experiments; the arguments just given even make this impossible in principle. The pursuits in this direction, so interesting they are, must therefore be avoided if we want to safely reach the top of Motion Mountain.

There are many more twists to this story. One possibility is that space-time itself, even without matter, is a bath. This speculation will be shown to be correct later on and seems to allow speaking of the wavefunction of all matter. But then again, it turns out that time is *undefined* at the scales where space-time would be an effective bath; this means that the concept of state is not applicable there.

A lack of 'state' for the universe is a strong statement. It also implies a lack of initial conditions! The arguments are precisely the same. This is a tough result. We are so used to think that the universe has initial conditions that we never question the term. (Even in this text the mistake might appear every now and then.) But there are no initial conditions of the universe.

We can retain as result, valid even in the light of the latest results of physics: the universe has *no* wavefunction and no initial conditions, independently of what is meant by 'universe'. But before we continue to explore the consequences of quantum theory for the whole universe, we study in more detail the consequences for our everyday observations.

26. Applied quantum mechanics – life, pleasure and the means to achieve them

Homo sum, humani nil a me alienum puto.*
Terence

Now that we can look at quantum effects without ideological baggage, let us have some serious fun in the world of quantum theory. The quantum of action has important consequences for biology, chemistry, technology and science fiction. We will only explore a cross section of these topics, but it will be worth it.

Biology

A special form of electromagnetic motion is of importance to humans: life. We mentioned at the start of quantum theory that life cannot be described by classical physics. Life is a quantum effect. Let us see why.

Living beings can be described as objects showing metabolism, information processing, information exchange, reproduction and motion. Obviously, all these properties follow from a single one, to which the others are enabling means:

▷ *Living beings are objects able to reproduce.**

From your biology lessons you might remember the some properties of heredity. Reproduction is characterized by random changes from one generation to the next. The statistics of mutations, for example Mendel's 'laws' of heredity, and the lack of intermediate states, are direct consequences of quantum theory. In other words, reproduction and growth are quantum effects.

In order to reproduce, living beings must be able to move in self-directed ways. An object able to perform self-directed motion is called a *machine*. All self-reproducing beings are machines.

Since reproduction and growth is simpler the smaller the system is, most living beings are extremely small machines for the tasks they perform, especially when compared to human made machines. This is the case even though the design of human machines has considerably fewer requirements: human-built machines do not need to be able to reproduce; as a result, they do not need to be made of a single piece of matter, as all living beings have to. But despite all the restrictions nature has to live with, living beings hold many miniaturization world records:

- The brain has the highest processing power per volume of any calculating device so far. Just look at the size of chess champion Gary Kasparov and the size of the computer against which he played.
- The brain has the densest and fastest memory of any device so far. The set of compact disks (CDs) or digital versatile disks (DVDs) that compare with the brain is many thousand times larger. Page 592

* 'I am a man and nothing human is alien to me.' Terence is Publius Terentius Afer (*c.* 190–159 BCE), the important roman poet. He writes this in his play Heauton Timorumenos, verse 77.

* However, there are examples of objects which reproduce and which nobody would call living. Can you find some examples, together with a sharper definition? Challenge 1238 n

▪ Motors in living beings are many orders of magnitude smaller than human-built ones. Just think about the muscles in the legs of an ant.

▪ The motion of living beings beats the acceleration of any human-built machine by orders of magnitude. No machine moves like a grasshopper.

▪ Living being's sensor performance, such as that of the eye or the ear, has been surpassed by human machines only recently. For the nose this feat is still far away. Nevertheless, the sensor sizes developed by evolution – think also about the ears or eyes of a common fly – is still unbeaten.

▪ Living beings that fly, swim or crawl – such as fruit flies, plankton or amoebas – are still thousands of times smaller than anything built by humans.

Challenge 1239 n

▪ Can you spot more examples?

The superior miniaturization of living beings is due to their continuous strife for efficient construction. In the structure of living beings, everything is connected to everything: each part influences many others. Indeed, the four basic processes in life, namely metabolic, mechanical, hormonal and electrical, are intertwined in space and time. For example, breathing helps digestion; head movements pump liquid through the spine; a single hormone influences many chemical processes. Furthermore, all parts in living systems have more than one function. For example, bones provide structure and produce blood; fingernails are tools and shed chemical waste.

The miniaturization, the reproduction, the growth and the functioning of living beings all rely on the quantum of action. Let us see how.

Reproduction

Life is a sexually transmitted disease.

Anonymous

All the astonishing complexity of life is geared towards reproduction. *Reproduction* is the ability of an object to build other objects similar to itself. Quantum theory told us that only a *similar* object is possible, as an exact copy would contradict the quantum of action, as we found out above.

Page 726

Since reproduction requires mass increase, reproducing objects show both metabolism and growth. In order that growth leads to an object similar to the original, a construction plan is necessary. This plan must be similar to the plan used by the previous generation. Organizing growth with a construction plan is only possible if nature is made of smallest entities which can be assembled following that plan.

We can thus deduce that reproduction implies that matter is made of smallest entities. If matter were not made of smallest entities, there would be no way to realize reproduction. Reproduction thus requires quantum theory. Indeed, without the quantum of action there would be no DNA molecules and there would be no way to inherit our own properties – our own construction plan – to children.

Passing on a plan requires that living beings have ways to store information. Living beings must have some built-in memory. We know already that a system with memory must be made of many particles. There is no other way to store information. The large number of particles is mainly necessary to protect the information from the influences of the outside world.

Our own construction plan, made of what biologists call genes, is stored in DNA molecules. Reproduction is thus first of all a transfer of parent's genes to the next generation. Page 782 We will come back to the details below. We first have a look on how our body moves itself and its genes around.

Quantum machines

Living beings are machines. How do these machines work? From a physical point of view, we need only a few sections of our walk so far to describe them: universal gravity and QED. Simply stated, life is an electromagnetic process taking place in weak gravity.* But the details of this statement are tricky and interesting. Table 57 gives an overview of motion processes in living beings. Interestingly, all motion in living beings can be summarized in a few classes by asking for the motor driving it.

Nature only needs few small but powerful devices to realize all motion types used by living beings. Given the long time that living systems have been around, these devices are extremely efficient. In fact, ion pumps, chemical pumps, rotational and linear molecular motors are all specialized molecular motors. Ion and chemical pumps are found in membranes and transport matter. Rotational and linear motor move structures against membranes. In short, all motion in living beings is due to molecular motors. Even though there is still a lot to be learned about them, what is known already is spectacular enough.

How do we move? – Molecular motors

How do our muscles work? What is the underlying motor? One of the beautiful results of modern biology is the elucidation of this issue. It turns out that muscles work because they contain molecules which change shape when supplied with energy. This shape change is repeatable. A clever combination and repetition of these molecular shape changes is then used to generate macroscopic motion. There are three basic classes of molecular motors: linear motors, rotational motors and pumps.

Linear motors are at the basis of muscle motion; other linear motors separate genes during cell division. They also move organelles inside cells and displace cells through the body during embryo growth, when wounds heal, or in other examples of cell motility. A typical molecular motor consumes around 100 to 1000 ATP molecules per second, thus about 10 to 100 aW. The numbers are small; however, we have to take into account that Challenge 1240 n the power white noise of the surrounding water is 10 nW. In other words, in every molecular motor, the power of the environmental noise is eight to nine orders of magnitude higher than the power consumed by the motor. The ratio shows what a fantastic piece of machinery such a motor is.

* In fact, also the nuclear interactions play some role for life: cosmic radiation is one source for random mutations, which are so important in evolution. Plant growers often use radioactive sources to increase Ref. 766 mutation rates. But obviously, radioactivity can also terminate life.

The nuclear interactions are also implicitly involved in several other ways. They were necessary to form the materials – carbon, oxygen, etc. – required for life. Nuclear interactions are behind the main mechanism for the burning of the Sun, which provides the energy for plants, for humans and for all other living beings (except a few bacteria in inaccessible places).

Summing up, the nuclear interactions play a role in the appearance and in the in destruction of life; but they play no (known) role for the actions of particular living beings.

Table 57 Motion and motors in living beings

Motion type	Examples	Main involved devices
Growth	collective molecular processes in cell growth	ion pumps
	gene turn-on and turn-off	linear molecular motors
	aging	linear molecular motors
Construction	materialtypes and properties (polysaccharides, lipids, proteins, nucleic acids, others)	material transport through muscles
	forces and interactions between biomolecules	cell membrane pumps
Functioning	details of metabolism (respiration, digestion)	muscles, ion pumps
	energy flow in biomolecules	
	thermodynamics of whole living system and of its parts	muscles
	muscle working	linear molecular motors
	nerve signalling	ion motion, ion pumps
	brain working	ion pumps
	illnesses	cell motility, chemical pumps
	viral infection of a cell	rotational molecular motors for RNA transport
Defence	the immune system	cell motility, linear molecular motors
Reproduction	information storage and retrieval	linear molecular motors inside cells, sometimes rotational motors, as in viruses
	cell division	linear molecular motors inside cells
	sperm motion	rotational molecular motors
	courting	muscles, brain, linear molecular motors
	evolution	muscles, linear molecular motors

Page 70 We encountered rotational motors already above. Nature uses them to rotate the cilia Ref. 769 of many bacteria as well as sperm tails. Researchers have also discovered that evolution produced molecular motors which turn around DNA helices like a motorized bolt would turn around a screw. Such motors are attached at the end of some viruses and insert the DNA into virus bodies when they are being built by infected cells, or extract the DNA Ref. 767 from the virus after it has infected a cell. Another rotational motor, the smallest known so far – 10 nm across and 8 nm high – is ATP synthase, a protein that synthesizes most ATP in cells.

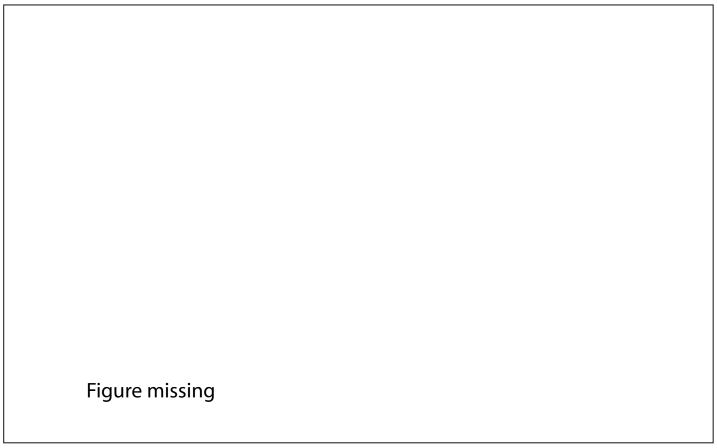


Figure 316 Myosin and actin: the building bricks of a simple linear molecular motor

The ways molecules produce movement in linear motors was uncovered during the Ref. 768 1990s. The results then started a wave of research on all other molecular motors found in nature. All molecular motors share a number of characteristic properties. There are no temperature gradients involved, as in car engines, no electrical currents, as in electrical motors, and no concentration gradients, as found in chemically induced motion. The central part of linear molecular motors is a combination of two protein molecules, namely myosin and actin. Myosin changes between two shapes and literally *walks* along actin. It moves in regular small steps. The motion step size has been measured with beautiful experiments to always be an integer multiple of 5.5 nm. A step, usually forward, but Ref. 768 sometimes backwards, results whenever an ATP (adenosine triphosphate) molecule, the standard biological fuel, hydrolyses to ADP (adenosine diphosphate), thus releasing its energy. The force generated is about 3 to 4 pN; the steps can be repeated several times a second. Muscle motion is the result of thousand of millions of such elementary steps taking place in concert.

How do molecular motors work? These motors are are so small that the noise due to the molecules of the liquid around them is not negligible. But nature is smart: with two tricks it takes advantage of Brownian motion and transforms it into macroscopic molecular motion. Molecular motors are therefore also called *Brownian motors*. The transformation of disordered molecular motion into ordered macroscopic motion is one of the great wonders of nature. The first trick of nature is the use of an asymmetric, but periodic potential, a so-called *ratchet*.* The second trick of nature is a temporal variation of the potential, together with an energy input to make it happen. The most important Ref. 770 realizations are shown in Figure 317.

The periodic potential variation allows that for a short time the Brownian motion of the moving molecule – typically 1 μm/s – affects its position. Then the molecule is fixed again. In most of these short times of free motion, the position will not change. But if the position does change, the intrinsic asymmetry of the ratchet shape ensures that in most cases the molecule advances in the preferred direction. Then the molecule is fixed again,

* It was named by Walt Disney after by Ratchet Gearloose, the famous inventor from Duckburg.

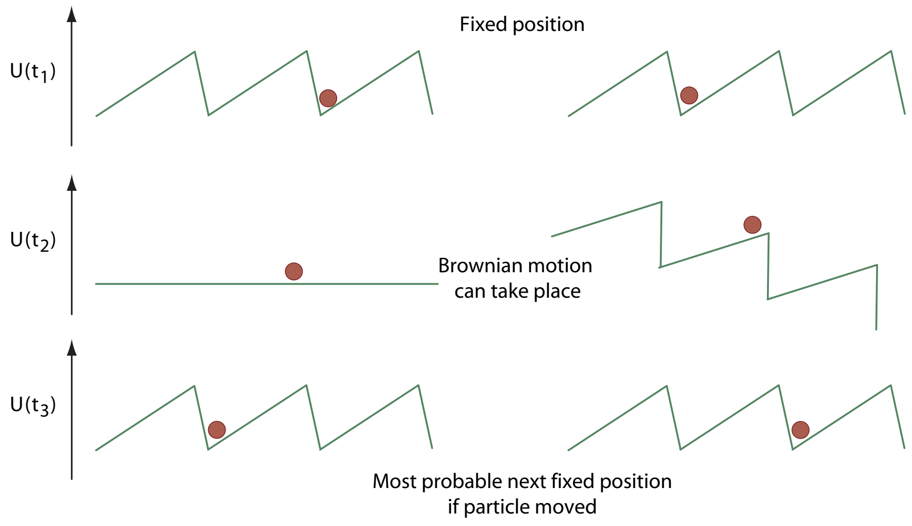


Figure 317 Two types of Brownian motors: switching potential (left) and tilting potential (right)

waiting for the next potential change. On average, the myosin molecule will thus move in one direction. Nowadays the motion of single molecules can be followed in special experimental set-ups. These experiments confirm that muscles use such a ratchet mechanism. The ATP molecule adds energy to the system and triggers the potential variation through the shape change it induces in the myosin molecule. That is how our muscles work.

Another well-studied linear molecular motor is the kinesin–microtubule system which carries organelles from one place to the other within a cell. As in the previous example, also in this case chemical energy is converted into unidirectional motion. Researchers were able to attach small silica beads to single molecules and to follow their motion. Using laser beams, they could even apply forces to these single molecules. Kinesin was found to move with around 800 nm/s, in steps lengths which are multiples of 8 nm, using one ATP molecule at a time, and exerting a force of about 6 pN.

Quantum ratchets also exist as human built systems, such as electrical ratchets for electron motion or optical ratchets that drive small particles. Extensive experimental research is going on in the field.

Curiosities and fun challenges about biology

The physics of life is still not fully explored.

Challenge 1241 n
- How would you determine which of two identical twins is the father of a baby?

Challenge 1242 n
- Can you give at least five arguments to show that a human clone, if there will ever be one, is a completely different person that the original? In fact, the first cloned cat, born in 2002, looked completely different from the 'original' (in fact, its mother). The fur colour and its patch pattern were completely different from that of the mother. Analogously, identical human twins have different finger prints, iris scans, blood vessel networks, intrauterine experiences, among others.
- Many molecules found in living beings, such as sugar, have mirror molecules. How-

Table 58 Approximate number of living species

Life group	Described species	Estimated species min.	Estimated species max.
Viruses	4 000	$50 \cdot 10^3$	$1 \cdot 10^6$
Prokaryotes ('bacteria')	4000	$50 \cdot 10^3$	$3 \cdot 10^6$
Fungi	72 000	$200 \cdot 10^3$	$2.7 \cdot 10^6$
Protozoa	40 000	$60 \cdot 10^3$	$200 \cdot 10^3$
Algae	40 000	$150 \cdot 10^3$	$1 \cdot 10^6$
Plants	270 000	$300 \cdot 10^3$	$500 \cdot 10^3$
Nematodes	25 000	$100 \cdot 10^3$	$1 \cdot 10^6$
Crustaceans	40 000	$75 \cdot 10^3$	$200 \cdot 10^3$
Arachnids	75 000	$300 \cdot 10^3$	$1 \cdot 10^6$
Insects	950 000	$2 \cdot 10^6$	$100 \cdot 10^6$
Molluscs	70 000	$100 \cdot 10^3$	$200 \cdot 10^3$
Vertebrates	45 000	$50 \cdot 10^3$	$55 \cdot 10^3$
Others	115 000	$200 \cdot 10^3$	$800 \cdot 10^3$
Total	$1.75 \cdot 10^6$	$3.6 \cdot 10^6$	$112 \cdot 10^6$

ever, in all living beings only one of the two sorts is found. Life is intrinsically asymmetric. How can this be? Challenge 1243 n

▪ How is it possible that the genetic difference between man and chimpanzee is regularly given as about 1 %, whereas the difference between man and woman is one chromosome in 46, in other words, about 2.2 %? Challenge 1244 n

▪ What is the longest time a single bacterium has survived? Not 5000 years as the bacteria found in Egyptian mummies, not even 25 million years as the bacteria resurrected from the intestines in insects enclosed in amber. It is much longer, namely an astonishing 250 million years. This is the time that bacteria discovered in the 1960s by Hans Pflug in (low-radioactivity) salt deposits in Fulda (Germany) have hibernated there before being brought back to life in the laboratory. The result has been recently confirmed by the discovery of another bacterium in a North-American salt deposit in the Salado formation.

▪ In 1967, a TV camera was deposited on the Moon. Unknown to everybody, it contained a small patch of *Streptococcus mitis*. Three years later, the camera was brought back to Earth. The bacteria were still alive. They had survived for three years without food, water or air. Life can be resilient indeed. Ref. 381

▪ In biology classifications are extremely useful. This is in full contrast to the situation in physics. Table 58 gives an overview of the magnitude of the task. This wealth of material can be summarized in one graph, shown in Figure 318. Ref. 771

Newer research seems to suggest some slight changes to the picture. So far however, there still is only a single root to the tree.

▪ How did life start? Challenge 1245 r

▪ Could life have arrived to Earth from outer space? Challenge 1246 n

▪ Life is not a clearly defined concept. The definition used above, the ability to repro-

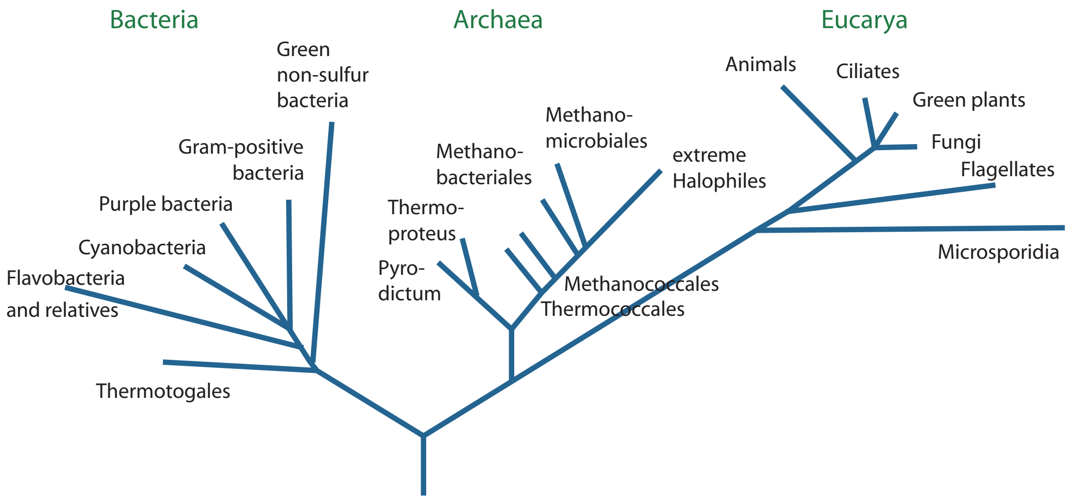


Figure 318 A modern version of the evolutionary tree

duce, has its limits when applied to old animals, to a hand cut off by mistake, to sperm or to ovules. It also gives problems when trying to apply it to single cells. Is the definition of life as 'self-determined motion in the service of reproduction' more appropriate? Or is
Challenge 1247 ny the definition of living beings as 'what is made of cells' more precise?

▪ Also growth is a type of motion. Some is extremely complex. Take the growth of acne. It requires a lack of zinc, a weak immune system, several bacteria, as well as the help of *Demodex brevis*, a mite (a small insect) that lives in skin pores. With a size of 0.3 mm, somewhat smaller than the full stop at the end of this sentence, this and other animals living on the human face can be observed with the help of a strong magnifying glass.

Humans have many living beings on board. For example, humans need bacteria to live. It is estimated that 90% of the bacteria in the human mouth alone are not known yet; only 500 species have been isolated so far. These useful bacteria help us as a defence against the more dangerous species.

▪ Mammals have a narrow operating temperature. In contrast to machines, humans
Challenge 1248 d function only if the internal temperature is within a narrow range. Why? Does this requirement also apply to extraterrestrials – provided they exist?

▪ How did the first cell arise? This question is still open. However, researchers have found several substances that spontaneously form closed membranes in water. Such substances also form foams. It might well be that life formed in foam.

The physics of pleasure

What is mind but motion in the intellectual sphere?
Oscar Wilde (1854–1900) *The Critic as Artist.*

Pleasure is a quantum effect. The reason is simple. Pleasure comes from the senses. All senses measure. And all measurements rely on quantum theory. The human body, like an

Figure 319 The different speed of the eye's colour sensors, the cones, lead to a strange effect when this picture (in colour version) is shaken right to left in *weak* light

expensive car, is full of sensors. Evolution has build these sensors in such a way that they trigger pleasure sensations whenever we do with our body what we are made for.

Of course, no scientist will admit that he studies pleasure. So he says that he studies the senses. The field is fascinating and still evolving; here we can only have a quick tour of the present knowledge.

Among the most astonishing aspects of our body sensors is their sensitivity. The *ear* is so sensitive and at the same time so robust against large signals that the experts are still studying how it works. No known sensor can cover an energy range of 10^{13}; the detected intensity ranges from $1\,\mathrm{pW/m^2}$ to $10\,\mathrm{W/m^2}$, the corresponding air pressure variations from $20\,\mu\mathrm{Pa}$ to 60 Pa.

Audible sound wavelengths span from 17 m (20 Hz) to 17 mm (20 kHz). In this range, the ear is able to distinguish at least 1500 pitches with its 16 000 to 20 000 hair cells. But the ear is also able to distinguish 400 from 401 Hz using a special pitch sharpening mechanism.

The *eye* is a position dependent photon detector. Each eye contains around 126 million separate detectors on the retina. Their spatial density is the highest possible that makes sense, given the diameter of the lens of the eye. They give the eye a resolving power of $1'$ and the capacity to detect down to 60 incident photons in 0.15 s, or 4 absorbed photons in the same time. There are about 6 million less sensitive colour detectors, the cones, whose distribution we have seen earlier on. The different chemicals in the three cone types (red, green, blue) lead to different sensor speeds; this can be checked with the simple test shown in Figure 319. The images of the eye are only sharp if the eye constantly moves in small random motions. If this motion is stopped, for example with chemicals, the images produced by the eye become unsharp.

Page 542

Ref. 772

The eye also contains 120 million highly sensitive general light intensity detectors, the rods. This sensitivity difference is the reason that at night all cats are grey. Until recently, human built light sensors with the same sensitivity as rods had to be helium cooled, because technology was not able to build sensors at room temperature as sensitive as the human eye.

The *touch sensors* are distributed over the skin, with a surface density which varies from one region to the other. It is lowest on the back and highest in the face and on the tongue. There are separate sensors for pressure, for deformation, for vibration, for tickling, for heat, for coldness, and for pain. Some react proportionally to the stimulus intensity, some differentially, giving signals only when the stimulus changes.

The taste mechanisms of *tongue* are only partially known. The tongue produces five taste signals* – sweet, salty, bitter, sour, proteic – and the mechanisms are just being unravelled. Democritus imagined that taste depends on the shape of atoms. Today it is known that sweet taste is connected with certain shape of molecules. Despite all this, no sensor with a distinguishing ability of the same degree as the tongue has yet been built by humans.

The *nose* has about 350 different smell receptors; through combinations it is estimated Ref. 773 that it can smell about 10 000 different smells. Together with the five signals that the sense of taste can produce, the nose also produces a vast range of taste sensations. It protects against chemical poisons, such as smoke, and against biological poisons, such as faecal matter. In contrast, artificial gas sensors exist only for a small range of gases. Good artificial taste and smell sensors would allow to check wine or cheese during their production, Challenge 1250 r thus making its inventor extremely rich.

The human body also contains orientation sensors in the ear, extension sensors in each muscle, pain sensors almost all over the skin and inside the body, heat sensors and coldness sensors on the skin and in other places. Other animals feature additional types of Page 489 sensors. Sharks can feel electrical fields, snakes have sensors for infrared; both are used to locate prey. Pigeons, trout and sharks can feel magnetic fields, and use this sense for navigation. Many birds and certain insects can see UV light. Bats are able to hear ultrasound up to 100 kHz and more. Whales and elephants can detect and localize infrasound signals.

In summary, the sensors with which nature provides us are state of the art; their sensitivity and ease of use is the highest possible. Since all sensors trigger pleasure or help to Ref. 774 avoid pain, nature obviously wants us to enjoy life with the most intense pleasure possible. Studying physics is one way to do this.

> There are two things that make life worth living: Mozart and quantum mechanics.
>
> Victor Weisskopf**

* Taste sensitivity is *not* separated on the tongue into distinct regions; this is an incorrect idea that has been copied from book to book for over a hundred years. You can perform a falsification by yourself, using sugar Challenge 1249 n or salt grains.

** Victor Friedrich Weisskopf (b. 1908 Vienna, d. 2002 Cambridge), acclaimed theoretical physicist who worked with Einstein, Born, Bohr, Schrödinger and Pauli. He catalysed the development of quantum electrodynamics and nuclear physics. He worked on the Manhattan project but later in life intensely campaigned against the use of nuclear weapons. During the cold war he accepted the membership in the Soviet Academy of Sciences. He was professor at MIT and for many years director of CERN, in Geneva. He wrote several successful physics textbooks. The author heard him making the above statement in Geneva, in 1981, during one of his lectures.

The nerves and the brain

> There is no such thing as perpetual tranquillity of mind while we live here; because life itself is but motion, and can never be without desire, nor without fear, no more than without sense.
>
> Thomas Hobbes (1588–1679) *Leviathan.*

The main unit processing all these signals, the brain, is another of the great wonders of nature. The human brain has the highest complexity of all brains known;* its processing power and speed is orders of magnitude larger than any device build by man.

Page 195 We saw already how electrical signals from the sensors are transported into the brain. In the brain, the arriving signals are classified and stored, sometimes for a short time, sometimes for a long time. The details of the various storage mechanisms, essentially taking place in the structure and the connection strength between brain cells, were elucidated by modern neuroscience. The remaining issue is the process of classification. For certain low level classifications, such as colours or geometrical shapes for the eye or sound harmonies for the ear, the mechanisms are known. But for high-level classifications, such as the ones used in conceptual thinking, the aim is not yet achieved. It is not well known how to describe the processes of reading, understanding and talking in terms of signal motions. Research is still in full swing and will remain so for the largest part of the twenty-first century.

In the following we have a look at a few abilities of our brain, of our body and of other bodies which are important for the study of motion.

Clocks in quantum mechanics

> L'horologe fait de la réclame pour le temps.*
>
> Georges Perros

Challenge 1251 n Most clocks used in everyday life are electromagnetic. (Do you know an exception?) Any clock on the wall, be it mechanical, quartz controlled, radio or solar controlled, or of any other type, is based on electromagnetic effects. There are even clocks of which we do not even know how they work. Just look at singing. We know from everyday experience that Ref. 775 humans are able to keep the beat to within a few per cent for a long time. Also when we sing a musical note we reproduce the original frequency with high accuracy. In many movements humans are able to keep time to high accuracy, e.g. when doing sport or when dancing. (For shorter or longer times, the internal clocks are not so precise.)

In addition, all clocks are limited by quantum mechanics, including the simple pendulum. Let us explore the topic.

* This is not in contrast with the fact that one or two whale species have brains with a slightly larger mass. The larger mass is due to the protection these brains require against the high pressures which appear when whales dive (some dive to depths of 1 km). The number of neurons in whale brains is considerably smaller than in human brains.

* Clocks are ads for time.

Do clocks exist?

Die Zukunft war früher auch besser.*
Karl Valentin, German writer.

In general relativity, we found that clocks do not exist, because there is no unit of time that can be formed using the constants c and G. Clocks, like any measurement standard, need matter and non-gravitational interactions to work.. This is the domain of quantum theory. Let us see what the situation is in this case.

Page 455

First of all, the time operator, or any operator proportional to it, is not an observable. Indeed, the time operator is not Hermitean, as any observable must be. In other words, there is no physical observable whose value is proportional to time. On the other hand, clocks are quite common; for example, the Sun or Big Ben work to everybody's satisfaction. Nature thus encourages us to look for an operator describing the position of the hands of a clock. However, if we look for such an operator we find a strange result. Any quantum system having a Hamiltonian bounded from below – having a lowest energy – lacks a Hermitean operator whose expectation value increases monotonically with time. This result can be proven rigorously. In other words, quantum theory states that time cannot be measured.

Challenge 1252 d

That time cannot be measured is not really a surprise. The meaning of this statement is that every clock needs to be wound up after a while. Take a mechanical pendulum clock. Only if the weight driving it can fall forever, without reaching a bottom position, can the clock go on working. However, in all clocks the weight has to stop when the chain end is reached or when the battery is empty. In other words, in all real clocks the Hamiltonian is bounded from below.

In short, quantum theory says that any clock can only be *approximate*. Quantum theory shows that exact *clocks do not exist in nature*. Obviously, this result can be of importance only for high precision clocks. What happens if we try to increase the precision of a clock as much as possible?

High precision implies high sensitivity to fluctuations. Now, all clocks have a motor inside that makes them work. A high precision clock thus needs a high precision motor. In all clocks, the position of the motor is read out and shown on the dial. The quantum of action implies that a precise clock motor has a position indeterminacy. The clock precision is thus limited. Worse, like any quantum system, the motor has a small, but finite probability to stop or to run backwards for a while.

You can check this prediction yourself. Just have a look at a clock when its battery is almost empty, or when the weight driving the pendulum has almost reached the bottom position. It will start doing funny things, like going backwards a bit or jumping back and forward. When the clock works normally, this behaviour is only strongly reduced in amount; however, it is still possible, though with low probability. This is true even for a sundial.

Challenge 1253 ny

In other words, clocks necessarily have to be macroscopic in order to work properly. A clock must be large as possible, in order to average out its fluctuations. Astronomical systems are examples. A good clock must also be well-isolated from the environment,

* Also the future used to be better in the past.

such as a freely flying object whose coordinate is used as time variable, as is done in certain optical clocks.

How big is the problem we have thus discovered? What is the actual error we make when using clocks? Given the various limitations due to quantum theory, what is the ultimate precision of a clock?

To start with, the indeterminacy relation provides the limit that the mass M of a clock must be larger than

Challenge 1254 ny

$$M > \frac{\hbar}{c^2 \tau} \tag{567}$$

Challenge 1255 e

which is obviously always fulfilled in everyday life. But we can do better. Like for a pendulum, we can relate the accuracy τ of the clock to its maximum reading time T. The idea was first published by Salecker and Wigner. They argued that

Ref. 776

$$M > \frac{\hbar}{c^2 \tau} \frac{T}{\tau} \tag{568}$$

where T is the time to be measured. You might check that this directly requires that any clock must be *macroscopic*.

Challenge 1256 e

Let us play with the formula by Salecker and Wigner. One way to rephrase it is the following. They showed that for a clock which can measure a time t, the size l is connected to the mass m by

$$l > \sqrt{\frac{\hbar t}{m}} \ . \tag{569}$$

Ref. 778

How close can this limit be achieved? It turns out that the smallest clocks known, as well as the clocks with most closely approach this limit are bacteria. The smallest bacteria, the *mycoplasmas*, have a mass of about $8 \cdot 10^{-17}$ kg, and reproduce every 100 min, with a precision of about 1 min. The size predicted from expression (569) is between 0.09 µm and 0.009 µm. The observed size of the smallest mycoplasmas is 0.3 µm. The fact that bacteria can come so close to the clock limit shows us again what a good engineer evolution has been.

Note that the requirement by Salecker and Wigner is not in contrast with the possibility to make the *oscillator* of the clock very small; people have built oscillators made of a single atom. In fact, such oscillations promise to be the most precise human built clocks.

Ref. 777

In the real world, the expression can be stated even more strictly. The whole mass M cannot be used in the above limit. For clocks made of atoms, only the binding energy between atoms can be used. This leads to the so-called *standard quantum limit for clocks*; it limits their frequency ν by

$$\frac{\delta \nu}{\nu} = \sqrt{\frac{\Delta E}{E_{\text{tot}}}} \tag{570}$$

where $\Delta E = \hbar / T$ is the energy indeterminacy stemming from the finite measuring time T and $E_{\text{tot}} = N E_{\text{bind}}$ is the total binding energy of the atoms in the meter bar. However, the quantum limit has not been achieved for clocks, even though experiments are getting near to it.

In summary, clocks exist only in the limit of $\hbar$ being negligible. In practice, the errors made by using clocks and meter bars can be made as small as required; it suffices to make the clocks large enough. We can thus continue our investigation into the details of matter without much worry. Only in the third part of our mountain ascent, where the precision requirements will be higher and general relativity will limit the size of physical systems, things will get much more interesting: the impossibility to build clocks will then become a central issue.

Living clocks

Among many things, living beings process information. Also computers do this, and like computers, all living beings need a clock to work well. Every clock needs is made up of the same components. It needs an *oscillator* determining the rhythm and a mechanism to feed the oscillator with energy. A clock also needs an oscillation *counter*, i.e. a mechanism that reads out the clock signal; a means of *signal distribution* throughout the system is required, synchronizing the processes attached to it. In addition, a clock needs a *reset mechanism*. If the clock has to cover many time scales, it needs several oscillators with different oscillation frequencies and a way to reset their relative phases.

Even though physicists know the details of technical clock building fairly well, we still do not know many parts of biological clocks. Most oscillators are chemical systems; Ref. 780 some, like the heart muscle or the timers in the brain, are electrical systems. The general elucidation of chemical oscillators is due to Ilya Prigogine; it has earned him a Nobel prize for chemistry in 1977. But not all the chemical oscillators in the human body are known yet, not to speak of the counter mechanisms. For example, a 24–minute cycle inside each human cell has been discovered only in 2003, and the oscillation mechanism is not yet fully clear. (It is known that a cell fed with heavy water ticks with 27–minute instead of Ref. 779 24–minute rhythm.) It might be that the daily rhythm, the circadian clock, is made up of or reset by 60 of these 24–minute cycles, triggered by some master cells in the human body. The clock reset mechanism for the circadian clock is also known to be triggered by daylight; the cells in the eye who perform this have been pinpointed only in 2002. The light signal is processed by the superchiasmatic nucleus, two dedicated structures in the brain's hypothalamus. The various cells in the human body react differently depending on the phase of this clock.

The clock with the longest cycle in the human body controls aging. A central mechanism for this clock seems to be the number of certain molecules attached to the DNA of the human chromosomes. At every division, one molecule is lost. When the molecules are all lost, the cell dies. Research into the mechanisms and the exceptions to this process (cancer cells, sexual cells) is ongoing.

The basis of the monthly period in women is equally interesting and complex.

The most fascinating clocks are those at the basis of conscious time. Of these, the brain's stopwatch or *interval timer*, has been most intensely studied. Only recently was its mechanism uncovered by combining data on human illnesses, human lesions, mag-Ref. 781 netic resonance studies, and effects of specific drugs. The basic mechanism takes place in the striatum in the basal ganglia of the brain. The striatum contains thousands of timer cells with different periods. They can be triggered by a 'start' signal. Due to their large number, for small times of the order of one second, every time interval has a different pattern across these cells. The brain can read these patterns and learn them. In this way we can time music or specific tasks to be performed, for example, one second after a signal.

Metre sticks

For length measurements, the situations is similar to that for time measurements. The limit by Salecker and Wigner can also be rewritten for length measurement devices. Are you able to do it?

Challenge 1257 ny

In general relativity we found that we need matter for any length measurement. Quantum theory, our description of matter, again shows that metre sticks are only approximately possible, but with errors which are negligible if the device is macroscopic.

Why are predictions so difficult, especially of the future?

> Future: that period of time in which our affairs prosper, our friends are true, and our happpiness is assured.
>
> Ambrose Bierce

If due to the quantum of action perfect clocks do not exist, is determinism still the correct description of nature?

We have seen that predictions of the future are made difficult by nonlinearities and the divergence of from similar conditions; we have seen that many particles make it difficult to predict the future due to the statistical nature of their initial conditions; we have seen that quantum theory makes it often hard to fully determine initial states; we have seen that not-trivial space-time topology can limit predictability; finally, we will discover that black hole and similar horizons can limit predictability due to their one-way transmission of energy, mass and signals.

Nevertheless, we also learned that all these limitations can be overcome for limited time intervals; in practice, these time intervals can be made so large that the limitations do not play a role in everyday life. In summary, in quantum theory both determinism and time remain applicable, as long as we do not extend it to infinite space and time. When extremely large dimensions and intervals need to be taken into account, quantum theory cannot be applied alone; in those cases, general relativity needs to be taken into account.

Decay and the golden rule

> I prefer most of all to remember the future.
>
> Salvador Dalì

The decoherence of superposition of macroscopically distinct states plays an important role in another common process: the decay of unstable systems or particles. Decay is any spontaneous change. Like the wave aspect of matter, decay is a process with no classical counterpart. True, decay, including the aging of humans, can be followed in classical physics; however, its *origin* is a pure quantum effect.

Experiments show that the prediction of decay, like that of scattering of particles, is only possible on average, for a large number of particles, never for a single one. These results confirm the quantum origin of the process. In every decay process, the superposition of macroscopically distinct states, which in this case are those of a decayed and an undecayed particle, is made to decohere rapidly by the interaction with the environment, and the 'environment' vacuum is sufficient to induce the decoherence. As usual, the details of the states involved are unknown for a single system and make any prediction for a single system impossible.

Decay is influenced by the environment, even in the case that it is 'only' the vacuum. This is true for all systems, including radioactive nuclei. The statement can be confirmed by experiment. By enclosing a part of space between two conducting plates, one can change the degrees of freedom of the vacuum contained between them. Putting an electromagnetically unstable particle, such as an excited atom, between the plates, indeed
Ref. 782 changes the lifetime of the particle. Can you explain why this method is not useful to
Challenge 1258 ny lengthen the lifespan of humans?

What is the *origin* of decay? Decay is always due to tunnelling. With the language of quantum electrodynamics, we can rephrase the answer: decay is motion induced by the vacuum fluctuations. The experiment between the plates confirms the importance of the environment fluctuations for the decay process.

For a system consisting of a large number N of identical particles, the decay is described by

$$\dot{N} = -\frac{N}{\tau} \quad \text{where}$$
$$\frac{1}{\tau} = \frac{2\pi}{\hbar} \left|\langle\psi_{\text{final}}|H_{\text{int}}|\psi_{\text{final}}\rangle\right|^2 \ . \qquad (571)$$

The decay is thus essentially an exponential one, independently of the details of the physical process. In addition, the decay time τ depends on the interaction and on the square modulus of the transition matrix element. This result was named the *golden rule* by Fermi,* because it works so well despite being an approximation whose domain of applicability is not easy to specify.

In practice, decay follows an exponential law. Experiments failed to see a deviation from this behaviour for over half a century. On the other hand, quantum theory shows
Ref. 783 that decay is exponential only in certain special systems. A calculation that takes into account higher order terms predicts deviations from exponential decay for completely isolated systems: for short times, the decay rate should vanish; for long times, the decay rate should follow an algebraic – not an exponential – dependence on time, in some
Ref. 784 cases even with superimposed oscillations. Only after an intense experimental search deviations for short times have finally been observed. The observation of deviations at long times are rendered impossible by the ubiquity of thermal noise. Theory shows that the exponential decay so regularly found in nature results only when the environment is noisy, the system made of many particles, or both. Since this is usually the case, the exceptional exponential decay becomes the (golden) rule in usual observations.

* Originally, the golden rule is an expression from the christian bible, namely the sentence 'Do to others what you want them to do to you'.

Zeno and the present in quantum theory

Utere tempore.*

Ovidius

As shown by perception research, what humans call 'present' has a duration of a few tenths Ref. 785 of a second. This leads us to ask whether the *physical* present might have a duration as well.

Every observation, like every photograph, implies a *time average*: observations average interactions over a given time. For a photograph, the duration is given by the shutter time; for a measurement, the average is defined by the set-up used. Whatever this set-up might be, the averaging time is never zero.

We thus need to ask whether the result of an observation will change if the observation time is shortened as much as possible, or if the observations will simply approach some limit situation. In everyday life, we are used to imagine that shortening the time taken to measure the position of a point object as much as possible will approach the ideal of a particle fixed at a given point in space. When Zeno discussed flight of an arrow, he assumed that this is possible.

Quantum theory has brought us so many surprises that the question should be studied carefully. We already know that the quantum of action makes rest an impossibility. However, the issue here is different: we are asking whether we can say that a system is at a given spot at a given time. In order to determine this, we could use a photographic camera whose shutter time can be reduced at will. What would we find? When the shutter time approaches the oscillation period of light, the sharpness of the image would decrease; in addition, the colour of the light would be influenced by the shutter motion. We can increase the energy of the light used, but the smaller wavelengths only shift the problem. At extremely small wavelengths, matter becomes transparent, and shutters cannot be realized any more. Quantum theory does not confirm the naive expectation that shorter shutter times lead to sharper images. In other words, the quantum aspects of the world show us that there is no way in principle to approach the limit that Zeno was discussing. Whenever one reduces shutter times as much as possible, observations become unsharp. This counter-intuitive result is due to the quantum of action: through the indeterminacy relation, the smallest action prevents that moving objects are at a fixed position at a given time. Zeno's discussion was based on an extrapolation of classical physics into domains where it is not valid any more. There is no 'point-like' instant of time that describes the present. The *present* is always an average over a non-vanishing interval of time.

What is motion?

Zeno was thus wrong in assuming that motion is a sequence of specific positions in space. Quantum theory implies that motion is *not* the change of position with time. The investigation of the issue showed that this statement is only an approximation for low energies or for long observation times.

How then can we describe motion in quantum theory? Quantum theory shows that motion is the low energy approximation of quantum evolution. Quantum evolution assumes that space and time measurements of sufficient precision can be performed. We

* 'Use your time.' *Tristia* 4, 3, 83

know that for any given observation energy, we can build clocks and meter bars with much higher accuracy than required, so that quantum evolution is applicable in all cases. *Motion is an approximation of quantum evolution.*

Obviously, this pragmatic description of motion rests on the assumption that for any observation energy we can find a still higher energy used by the measurement instruments to define space and time. We deduce that if a highest energy would exist in nature, we would get into big trouble, as quantum theory would then break down. As long as energy has no limits, all problems are avoided, and motion remains a sequence of quantum observables or quantum states, whichever you prefer.

The assumption of energy without limit works extremely well; it lies at the basis of the whole second part of the mountain ascent, even though it is rather hidden. In the third and final part, we will discover that there indeed is a maximum energy in nature, so that we will need to change our approach. However, this energy value is so huge that it does not bother us at all at this point of our exploration. But it will do so later on.

Several sections, e.g. on time and on the quantum Zeno effect, will be added here

Consciousness: a result of the quantum of action

Page 644 Consciousness is our ability to observe what is going on in our mind. This activity, like any type of change, can itself be observed and studied. Obviously, consciousness takes place in the brain. If it were not, there would be no way to keep it connected with a given person. We know that each brain moves with over one million kilometres per hour through the cosmic background radiation; we also observe that consciousness moves along with it.

The brain is a quantum system; it is based on molecules and electrical currents. The changes in consciousness that appear when matter is taken away from the brain – in operations or accidents – or when currents are injected into the brain – in accidents, experiments or misguided treatments – have been described in great detail by the medical profession. Also the observed influence of chemicals on the brain – from alcohol to hard drugs – makes the same point. The brain is a quantum system.

Magnetic resonance imaging can detect which parts of the brain work when sensing, remembering or thinking. Not only is sight, noise and thought processed in the brain; we can follow the processing on computer screens. The other, more questionable experimental method, positron tomography, works by letting people swallow radioactive sugar. It confirms the findings on the location of thought and on its dependence on chemical fuel. In addition, we already know that memory depends on the particle nature of matter. All these observations depend on the quantum of action.

Not only the consciousness of others, also your own consciousness is a quantum process. Can you give some arguments? Challenge 1259 ny

In short, we know that thought and consciousness are examples of motion. We are thus in the same situation as material scientists were before quantum theory: they knew that electromagnetic fields influence matter, but they could not say how electromagnetism was involved in the build-up of matter. We know that consciousness is made from the signal propagation and signal processing in the brain; we know that consciousness is an electrochemical process. But we do not know yet the details of how the signals make up consciousness. Unravelling the workings of this fascinating quantum system is the aim of neurological science. This is one of the great challenges of twenty-first century science.

It is sometimes claimed that consciousness is not a physical process. Every expert of motion should be able to convincingly show the opposite, even though the details are not clear yet. Can you add arguments to the ones given here? Challenge 1260 ny

Why can we observe motion?

Studying nature can be one of the most intense pleasures of life. All pleasure is based on the ability to observe motion. Our human condition is central to this ability. In our adventure so far we found that we experience motion only because we are of finite size, only because we are made of a large but finite number of atoms, only because we have a finite but moderate temperature, only because we are a mixture of liquids and solids, only because we are electrically neutral, only because we are large compared to a black hole of our same mass, only because we are large compared to our quantum mechanical wavelength, only because we have a limited memory, only because our brain forces us to approximate space and time by continuous entities, and only because our brain cannot avoid describing nature as made of different parts. If any of these conditions were not fulfilled we would not observe motion; we would have no fun studying physics.

In addition, we saw that we have these abilities only because our forefathers lived on Earth, only because life evolved here, only because we live in a relatively quiet region of our galaxy, and only because the human species evolved long after than the big bang.

If any of these conditions were not fulfilled, or if we were not humans (or animals), motion would not exist. In many ways motion is thus an illusion, as Zeno of Elea had claimed. To say the least, the observation of motion is due to the limitations of the human condition. A complete description of motion and nature must take this connection into account. Before we do that, we explore a few details of this connection.

Curiosities and fun challenges about quantum experience

The fascination of quantum effects is still lasting, despite over 100 years of intense studies.

- Are ghost images in TV sets, often due to spurious reflections, examples of interference? Challenge 1261 n
- What happens when two monochromatic electrons overlap? Challenge 1262 r
- The sense of smell is quite complex. For example, the substance that smells most badly to humans is skatole or 3-methylindole. This is the molecule to which the human nose is most sensitive. Skatole makes faeces smell bad; it is a result of hemoglobin entering the digestive tract through the bile. In contrast to humans, skatole attracts flies; it is also used by some plants for the same reason.

On the other hand, small levels of skatole do not smell bad to humans. It is also used

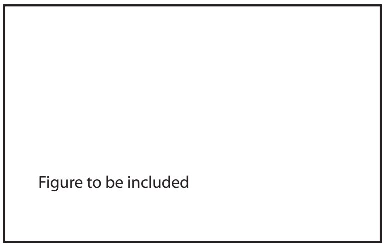


Figure 320 Atoms and dangling bonds

by the food industry in small quantities to give smell and taste to vanilla ice cream.

▪ It is worth noting that human senses detect energies of quite different magnitudes. The eyes can detect light energies of about 1 aJ, whereas the sense of touch can detect only
Challenge 1263 n energies as large as about 10 μJ. Is one of the two systems relativistic?

▪ Compared to all primates, the human eye is special: it is white, thus allowing others
Ref. 786 to see the direction in which one looks. Comparison with primates shows that the white colour has evolved to allow more communication between individuals.

▪ There is ample evidence that not using the senses is damaging. People have studied what happens when in the first years of life the vestibular sense – the one used for motion detection and balance restoration – is not used enough. Lack of rocking is extremely hard to compensate later in life. Also dangerous is the lack of use of the sense of touch. Babies, like all small mammals, that are generally and systematically deprived of these
Ref. 787 experiences tend to violent behaviour during the rest of their life.

Nature has indeed invented pleasure as a guide for a human behaviour. All of biology builds on chemistry and on materials science. Let's have a short overview of both fields.

Chemistry – from atoms to DNA

It is an old truth that Schrödinger's equation contains all of chemistry.* With quantum theory, for the first time people were able to calculate the strengths of chemical bonds, and what is more important, the angle between them. Quantum theory thus explains the *shape* of molecules and thus indirectly, the shape of all matter.

To understand molecules, the first step is to understand atoms. The early quantum theorists, above all Niels Bohr, spent a lot of energy in understanding their structure. The main result is that atoms are structured, though spherical electron clouds.

After more than thirty years of work by the brightest physicists in Göttingen and Copenhagen, it was found that electrons in atoms form various layers around the central nucleus. The layers are numbered from the inside by a number called the *principal quantum number*, usually written n.

Quantum theory shows that the first layer has room for two electrons, the second for 8, the third for 18 and the general n-th shell for $2n^2$ electrons. A way to picture this con-
Ref. 796 nection is shown in Figure 321. It is called the *periodic table of the elements*. The standard

* The precise statement is: the *Dirac* equation contains all of chemistry. The relativistic effects that
Ref. 788 distinguish the two equations are necessary, for example, to understand why gold is yellow and does not like to react or why mercury is liquid.

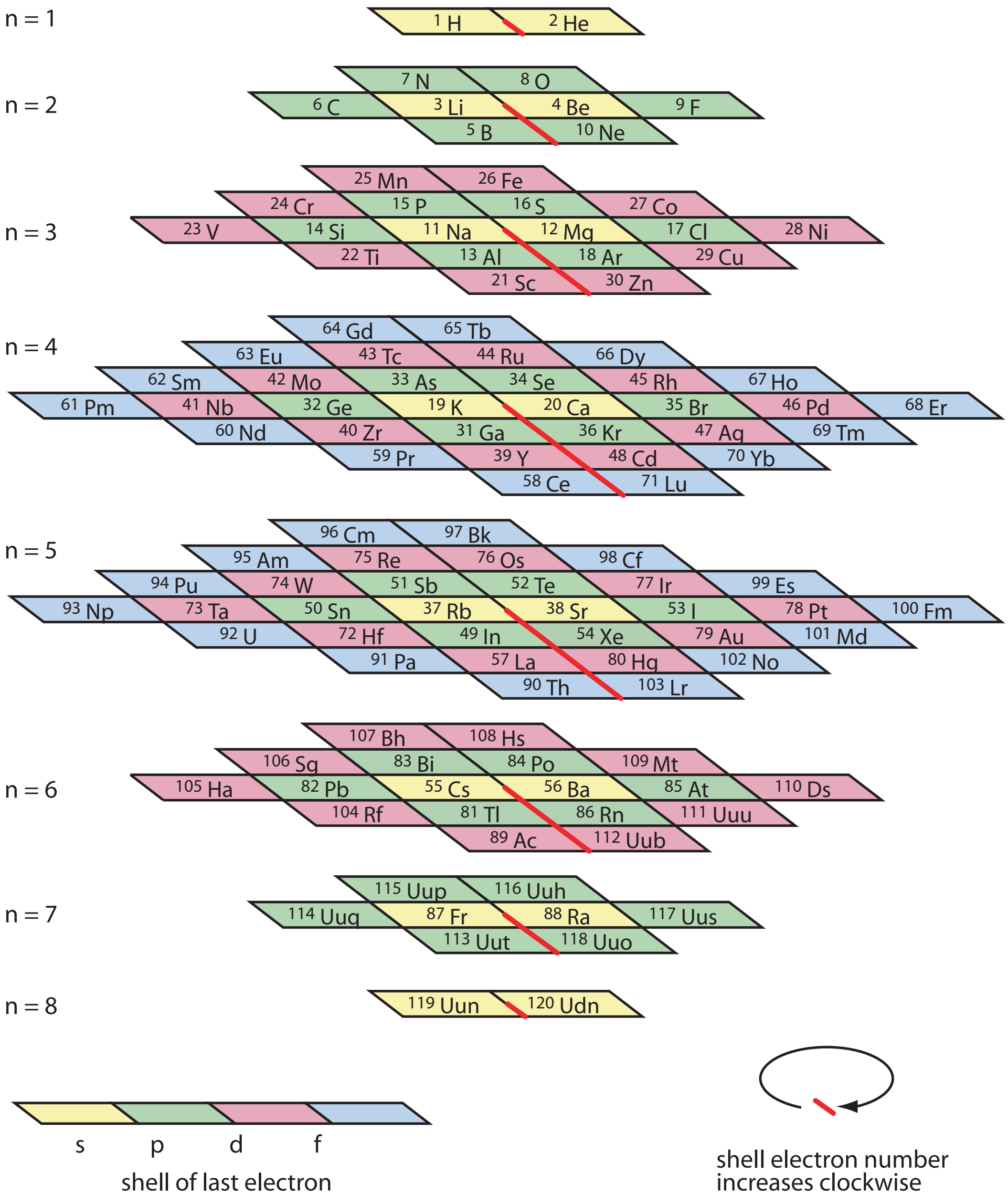


Figure 321 An unusual form of the periodic table of the elements

Page 1085 way to show the table is shown in Appendix C.

– CS – more to be added – CS –

When atoms approach each other, they can form one or several bonds. The preferred distance of these bonds, the angles between them, are due to the structure of the atomic

Figure missing

Figure 322 Several ways to picture dna

electron clouds.

Do you remember those funny pictures of school chemistry about orbitals and dangling bonds? Well, dangling bonds can now be seen. Several groups were able to image them using scanning force or scanning tunnelling microscopes. Ref. 797 Ref. 798

The angles between the bonds explain why the angle of tetrahedral skeletons ($2\arctan\sqrt{2} = 109.47°$) are so common in molecules. For example, the H-O-H angle in water molecules is 107°. Challenge 1264 e

At the centre of each atom cloud is the nucleus, which contains almost all the atomic mass. The nucleus consists of protons and neutrons. The structure of nuclei is even more complex than that of electron clouds. We explore it in a separate chapter later on. Page 836

Ribonucleic acid and Deoxyribonucleic acid

Probably the most fascinating molecule is human deoxyribonucleic acid. The nucleic acids where discovered in 1869 by the Swiss physician Friedrich Miescher (1844–1895) in white blood cells. In 1874 he published an important study showing that the molecule is contained in spermatozoa, and discusses the question if this substance could be related to heredity. With his work, Miescher paved the way to a research field that earned many colleagues Nobel Prizes (though not for himself).

DNA is a polymer, as shown in Figure 322, and is among the longest molecules known. Human DNA molecules, for example, can be up to 10 cm in length. It consists of a double helix of sugar derivates, to which four nuclei acids are attached in irregular order.

At the start of the twentieth century it became clear that Desoxyribonukleinsäure (DNS) or deoxyribonucleic acid (DNA) in English – was precisely what Erwin Schrödinger had predicted to exist in his book *What Is Life?* As central part of the chromosomes contained the cell nuclei, DNA is responsible for the storage and reproduction of the information on the construction and functioning of Eukariotes. The information is coded in the ordering of the four nucleic acids. DNA is the carrier of hereditary information. DNA determines in great part how the single cell we all once have been grows into the complex human machine we are as adults. For example, DNA determines the hair colour, predisposes for certain illnesses, determines the maximum size one can grow to, and much more. Of all known molecules, human DNA is thus most intimately related to human existence. The large size of the molecules is the reason that understanding its full structure and its full contents is a task that will occupy scientists for several generations to come.

Chemical challenges and curiosities

▪ Muscles produce motion through electrical stimulation. Can technical systems do the same? There is a candidate. So-called electroactive polymers change shape when they are activated with electrical current or with chemicals. They are lightweight, quiet and simple to manufacture. However, the first arm wrestling contest between human and artificial muscles held in 2005 was won by a teenage girl. The race to do better is ongoing.

▪ A cube of sugar does not burn. However, if you but some cigarette ash on top of it, it burns. Why?

Challenge 1265 ny ▪ Why do organic materials burn at much lower temperature than inorganic materials?

▪ An important aspect of life is death. When we die, conserved quantities like our energy, momentum, angular momentum and several other quantum numbers are redistributed. They are redistributed because conservation means that nothing is lost. What does all this imply for what happens after death? Challenge 1266 ny

Materials science

> Did you know that one cannot use a boiled egg as a toothpick?
>
> Karl Valentin

It was mentioned several times that the quantum of action explains all properties of matter. Many researchers from physics, chemistry, metallurgy, engineering, mathematics and biology have cooperated in the proof of this statement. In our mountain ascent we have little time to explore this vast topic. Let us walk through a selection.

Why does the floor not fall?

We do not fall through the mountain we are walking on. Some interaction keeps us from falling through. In turn, the continents keep the mountains from falling through them. Also the liquid magma in the Earth's interior keeps the continents from sinking. All these statements can be summarized. Atoms do not penetrate each other. Despite being mostly empty clouds, atoms keep a distance. All this is due to the Pauli principle between electrons. the fermion character of electrons avoids that atoms interpenetrate. At least on Earth. Page 732

Not all floors keep up due to the fermion character of electrons. Atoms are not impenetrable at all pressures. They can collapse, and form new types of floors. Some floors are so exciting to study that people have spent their whole life to understand why they do not fall, or when they do, how it happens: the surfaces of stars.

In most stars, the radiation pressure of the light plays only a minor role. Light pressure does play a role in determining the size of red giants, such as Betelgeuse; but for average stars, light pressure is negligible.

In most stars, such as in the Sun, the gas pressure takes the role which the incompressibility of solids and liquids has for planets. The pressure is due to the heat produced by the nuclear reactions.

The next star type appears whenever light pressure, gas pressure and the electronic Pauli pressure cannot keep atoms from interpenetrating. In that case, atoms are compressed until all electrons are pushed into the protons. Protons then become neutrons, and the whole star has the same mass density of atomic nuclei, namely about $2.3 \cdot 10^{17}$ kg/m^3. A drop weighs about 200 000 tons. In these so-called *neutron stars*, the floor – or better, the size – is also determined by Pauli pressure; however, it is the Pauli Page 849 pressure between neutrons, triggered by the nuclear interactions. These neutron stars are all around 10 km in radius.

If the pressure increases still further the star becomes a black hole, and never stops Page 441 collapsing. Black holes have no floor at all; they still have a constant size though, determined by the horizon curvature.

The question whether other star types exist in nature, with other floor forming mechanisms – such as quark stars – is still a topic of research.

Rocks and stones

If a geologist takes a stone his his hands, he is usually able to give, within an error of a few percent, the age of the stone. The full story forms a large part of geology, but the general lines should be known to every physicist.

Every stone arrives in your hand through the *rock cycle*. The rock cycle is a process that transforms magma from the interior of the Earth into *igneous rocks*, through cooling and crystallization. Igneous rocks, such as basalt, can transform through erosion, transport and deposition into *sedimentary rocks*. Either of these two rock types can be transformed through high pressures or temperatures into *metamorphic rocks*, such as marble. Finally, most rocks are generally – but not always – transformed back into magma.

The full rock cycle takes around 110 to 170 million years. For this reason, rocks that are older that this age much less common on Earth. Any stone is the product of erosion of one of the rock types. A geologist can usually tell, simply by looking at it, the type of rock it belongs to; if he sees the original environment, he can also give the age, without any laboratory.

How can one look through matter?

Quantum theory showed us that all obstacles have only finite potential heights. That leads to a question: Is it possible to look through matter? For example, can we see what is hidden inside a mountain? To be able to do this, we need a signal which fulfils two conditions: it must be able to *penetrate* the mountain, and it must be scattered in a *material-dependent* way. Table 60 gives an overview of the possibilities.

We see that many signals are able to penetrate a mountain. However, only sound or radio waves provide the possibility to distinguish different materials, or to distinguish solids from liquids and from air. In addition, any useful method requires a large number of signal sources and of signal receptors, and thus a large amount of cash. Will there ever be a simple method allowing to look into mountains as precisely as X-rays allow to study human bodies? For example, will it ever be possible to map the interior of the pyramids? Challenge 1267 n A motion expert like the reader should be able to give a definite answer.

One of the high points of twentieth century physics was the development of the best method so far to look into matter with dimensions of about a meter or less: magnetic Page 836 resonance imaging. We will discuss it later on.

Table 59 The types of rocks and stones

Type	Properties	Subtype	Example
Igneous rocks (magmatites)	formed from magma, 95% of all rocks	volcanic or extrusive	basalt (ocean floors, Giant's causeway), andesite, obsidian
		plutonic or intrusive	granite, gabbro
Sedimenary rocks (sedimentites)	often with fossils	clastic	shale, siltstone, sandstone
		biogenic	limestone, chalk, dolostone
		precipitate	halite, gypsum
Metamorphic rocks (metamorphites)	transformed by heat and pressure	foliated	slate, schist, gneiss (Himalayas)
		non-foliated (grandoblastic or hornfelsic)	marble, skarn, quartzite
Meteorites	from the solar system	rock meteorites	
		iron meteorites	

The other modern imaging technique, ultrasound imaging, is getting more and more criticized. It is much used for prenatal diagnostics of embryos. However, studies have found that ultrasound produces extremely high levels of audible sound to the baby, especially when the ultrasound is switched on or off, and that babies react negatively to this loud noise.

What is necessary to make matter invisible?

You might have already imagined what adventures would be possible if you could be invisible for a while. Some years ago, a team of Dutch scientists found a material than can be switched from mirror mode to transparent mode using an electrical signal. This seems a first step to realize the dream to become invisible at will. Ref. 789

Nature shows us how to be invisible. An object is invisible if it has no surface, no absorption and small size. In short, invisible objects are either small clouds or composed of them. Most atoms and molecules are examples. Homogeneous non-absorbing gases also realize these conditions. That is the reason that air is (usually) invisible. When air is not homogeneous, it can be visible, e.g. above hot surfaces.

In contrast to gases, solids or liquids do have surfaces. Surfaces are usually visible, even if the body is transparent, because the refractive index changes there. For example, quartz can be made so transparent that one can look through 1 000 km of it; pure quartz is thus more transparent than usual air. Still, objects made of pure quartz are visible to the eye due to the index change at the surface. Quartz can be invisible only when submerged in

Table 60 Signals penetrating mountains and other matter

Signal	Penetration depth in stone	Achieved resolution	Material dependence	Use
matter				
diffusion of water or liquid chemicals	*c.* 5 km	*c.* 100 m	medium	mapping hydrosystems
diffusion of gases	*c.* 5 km	*c.* 100 m	medium	studying vacuum systems
electromagnetism				
sound, explosions, seismic waves	0.1 – 10 m	*c.* $l/100$	high	oil and ore search, structure mapping in rocks
ultrasound		1 mm	high	medical imaging, acoustic microscopy
infrasound and earthquakes	100 000 km	100 km	high	mapping of Earth crust and mantle
static magnetic fields			medium	cable search, cable fault localization, search for structure inside soil and rocks via changes of the Earth's magnetic field
static electric fields		low	no use	
electrical currents				soil and rock investigations, search for tooth decay
radio waves	10 m	30 m to 1 mm	small	soil radar (up to 10 MW), magnetic imaging, research into solar interior
mm and THz waves	below 1 mm	1 mm		see through clothes, envelopes and teeth Ref. 790
infrared	*c.* 1 m	0.1 m	medium	mapping of soil over 100 m
visible light	*c.* 1 cm	0.1 µm	medium	imaging of many sorts
X rays	a few metre	5 µm	high	medicine, material analysis, airports
muons created by cosmic radiation	up to *c.* 300 m	0.1 m	small	finding caves in pyramids, imaging truck interiors
weak interactions				
neutrino beams	light years	zero	very weak	studies of Sun
strong interactions				
cosmic radiation	1 m to 1 km			
radioactivity	1 mm to 1 m			airports
gravitation				
change of gravitational acceleration		50 m	low	oil & ore search

Table 61 Matter at lowest temperatures

Phase	Type	Low temperature behaviour	Example
Solid	conductor	superconductivity	lead
	insulator	ferromagnetism	iron
	diamagnetism		
	antiferromagnet		
Liquid	bosonic	Bose–Einstein condensation, i.e. superfluidity	^{4}He
	fermionic	pairing, then BEC, i.e. superfluidity	^{3}He
Gas	bosonic	Bose–Einstein condensation	^{87}Rb, ^{7}Li, ^{23}Na, H, ^{4}He, ^{41}K
	fermionic	pairing, then Bose–Einstein condensation	^{40}K, ^{6}Li

liquids with the same refractive index.

In other words, to become invisible, we must transform ourselves into a diffuse cloud of non-absorbing atoms. On the way to become invisible, we would loose all memory and all genes, in short, we would loose all our individuality. But an individual cannot be made of gas. An individual is defined through its boundary. There is no way that we can be invisible; a reversible way to perform the feat is also impossible. In summary, quantum theory shows that only the dead can be invisible.

How does matter behave at lowest temperatures?

The low-temperature behaviour of matter has numerous experimental and theoretical aspects. The first issue is whether matter is always solid at low temperatures. The answer is no. All phases exist at low temperatures, as shown in Table 61.

Concerning the electric properties of matter at lowest temperatures, the present status is that matter is either insulating or superconducting. Finally, one can ask about the magnetic properties of matter at low temperatures. We know already that matter can not be paramagnetic at lowest temperatures. It seems that matter is either ferromagnetic, diamagnetic or antiferromagnetic at lowest temperatures.

Page 789

More about superfluidity and superconductivity will be told below.

Curiosities and fun challenges about materials science

Materials science is not a central part of this walk. A few curiosities can give a taste of it.

Ref. 791

Challenge 1268 ny

▪ What is the maximum height of a mountain? This question is of course of interest to all climbers. Many effects limit the height. The most important is the fact that under heavy pressure, solids become liquid. For example, on Earth this happens at about 27 km. This is quite a bit more than the highest mountain known, which is the volcano Mauna Kea in Hawaii, whose top is about 9.45 km above the base. On Mars gravity is weaker, so that mountains can be higher. Indeed the highest mountain on Mars, Olympus mons, is

Challenge 1269 n 80 km high. Can you find a few other effects limiting mountain height?

▪ Do you want to become rich? Just invent something that can be produced in the factory, is cheap and can substitute duck feathers in bed covers, sleeping bags or in badminton
Challenge 1270 r shuttlecocks. Another industrial challenge is to find an artificial substitute for latex, and a third one is to find a substitute for a material that is rapidly disappearing due to pollution: cork.

Challenge 1271 ny ▪ What is the difference between solids, liquids and gases?

▪ What is the difference between the makers of bronze age knifes and the builders of the Eiffel tower? Only their control of dislocation distributions.

▪ Quantum theory shows that tight walls do not exist. Every material is penetrable.
Challenge 1272 n Why?

▪ Quantum theory shows that even if tight walls would exist, the lid of such a box can
Challenge 1273 n never be tightly shut. Can you provide the argument?

▪ Quantum theory predicts that heat transport at a given temperature is quantized.
Challenge 1274 ny Can you guess the unit of thermal conductance?

▪ Robert Full has shown thatvan der Waals forces are responsible for the way that
Ref. 792 geckos walk on walls and ceilings. The gecko, a small reptile with a mass of about 100 g, uses an elaborate structure on its feet to perform the trick. Each foot has 500 000 hairs each split in up to 1000 small spatulae, and each spatula uses the van der Waals force (or alternatively, capillary forces) to stick to the surface. As a result, the gecko can walk on vertical glass walls or even on glass ceilings; the sticking force can be as high as 100 N per foot.

The same mechanism is used by jumping spiders (*Salticidae*). For example, *Evarcha arcuata* have hairs at their feet which are covered by hundred of thousands of setules.
Ref. 793 Again. the van der Waals force in each setule helps the spider to stick on surfaces.

Researchers have copied these mechanisms for the first time in 2003, using microlithography on polyimide, and hope to make durable sticky materials in the future.

▪ Millimetre waves or terahertz waves are emitted by all bodies at room temperature. Modern camera systems allow to image them. In this way, it is possible to see through clothes. This ability could be used in future to detect hidden weapons in airports. But the development of a practical and affordable detector which can be handled as easily as a binocular is still under way. The waves can also be used to see through paper, thus making it unnecessary to open letters in order to read them. Secret services are exploiting this technique. A third application of terahertz waves might be in medical diagnostic, for example for the search of tooth decay. Terahertz waves are almost without side effects, and thus superior to X-rays. The lack of low-priced quality sources is still an obstacle to their application.

▪ Does the melting point of water depend on the magnetic field? This surprising claim
Ref. 794 was made in 2004 by Inaba Hiadeaki and colleagues. They found a change of 0.9 mK/T. It is known that the refractive index and the near infrared spectrum of water is affected by magnetic fields. Indeed, not everything about water might be known yet.

▪ Plasmas, or ionized gases, are useful for many applications. Not only can they be used for heating or cooking and generated by chemical means (such plasmas are variously called fire or flames) but they can also be generated electrically and used for lighting or deposition of materials. Electrically generated plasmas are even being studied for the disinfection of dental cavities.

▪ It is known that the concentration of CO_2 in the atmosphere between 1800 and 2005 has increased from 280 to 380 parts per million. (How would you measure this?) It is known without doubt that this increase is due to human burning of fossil fuels, and not to natural sources such as the oceans or volcanos. There are three arguments. First of all, there was a parallel decline of the $^{14}C/^{12}C$ ratio. Second, there was a parallel decline of the $^{13}C/^{12}C$ ratio. Finally, there was a parallel decline of the oxygen concentration. All three measurements independently imply that the CO_2 increase is due to the burning of fuels, which are low in ^{14}C and in ^{13}C, and at the same time decrease the oxygen ratio. Natural sources do not have these three effects. Since CO_2 is a major greenhouse gas, the data implies that humans are also responsible for a large part of the temperature increase during the same period.

Challenge 1275 ny

▪ The technologies to produce perfect crystals, without grain boundaries or dislocations, are an important part of modern industry. Perfectly regular crystals are at the basis of the integrated circuits used in electronic appliances, are central to many laser and telecommunication systems and are used to produce synthetic jewels.

Synthetic diamonds have now displaced natural diamonds in almost all applications. In the last years, methods to produce large, white, jewel-quality diamonds of ten carats and more are being developed. These advances will lead to a big change in all the domains that depend on these stones, such as the production of the special chirurgical knives used in eye lense operation.

Ref. 795

▪ How can a small plant pierce through tarmac?

Challenge 1276 ny

Quantum technology

> I were better to be eaten to death with a rust
> than to be scoured to nothing with perpetual motion.
> William Shakespeare (1564–1616) *King Henry IV*.

Quantum effects do not appear only in microscopic systems. Several quantum effects are important in modern life; transistors, lasers, superconductivity and a few other effects and systems are worth knowing.

Motion without friction – superconductivity and superfluidity

We are used to think that friction is inevitable. We even learned that friction was an inevitable result of the particle structure of matter. I should come to the surprise of every physicist that motion without friction is possible.

In 1911 Gilles Holst and Heike Kamerlingh Onnes discovered that at low temperatures, electric currents can flow with no resistance, i.e., with no friction, through lead. The observation is called *superconductivity*. In the century after that, many metals, alloys and ceramics have been found to show the same behaviour.

The condition for the observation of motion without friction is that quantum effects play an essential role. That is the reason for the requirement of low temperature in such experiments. Nevertheless, it took over 50 years to reach a full understanding of the effect. This happened in 1957, when Bardeen, Cooper and Schrieffer published their results. At low temperatures, electron behaviour is dominated by an attractive interaction that

Ref. 805

Figure missing

Figure 323 Nuclear magnetic resonance shows that vortices in superfluid ^{3}He-B are quantized.

makes them form pairs – today called *Cooper pairs* – that are effective bosons. And bosons can all be in the same state, thus effectively moving without friction.

For superconductivity, the attractive interaction between electrons is due to the deformation of the lattice. Two electrons attract each other in the same way as two masses attract each other due to deformation of the space-time mattress. However, in the case of solids, the deformations are quantized. With this approach, Bardeen, Cooper and Schrieffer explained the lack of electric resistance of superconducting materials, their complete diamagnetism ($\mu_r = 0$), the existence of an energy gap, the second-order transition to normal conductivity at a specific temperature, and the dependence of this temperature on the mass of the isotopes. Last but not least, they received the Nobel prize in 1972. *

Another type of motion without friction is *superfluidity*. Already in 1937, Pyotr Kapitsa had predicted that normal helium (^{4}He), below a transition observed at the temperature of 2.17 K, would be a superfluid. In this domain, the fluid moves without friction through tubes. (In fact, the fluid remains a mixture of a superfluid component and a normal component.) Helium is even able, after an initial kick, to flow over obstacles, such as glass walls, or to flow out of bottles. ^{4}He is a boson, so no pairing is necessary for it to flow without friction. This research earned Kapitsa a Nobel prize in 1978.

The explanation of superconductivity also helped for fermionic superfluidity. In 1972, Richardson, Lee, and Osheroff found that even ^{3}He is superfluid at temperatures below 2.7 mK. ^{3}He is a fermion, and requires pairing to become superfluid. In fact, below 2.2 mK, ^{3}He is even superfluid in two different ways; one speaks of phase A and phase B.**

In this case, the theoreticians had been faster. The theory for superconductivity through pairing had been adapted to superfluids already in 1958 – before any data were available – by Bohr, Mottelson and Pines. This theory was then adapted again by Anthony Leggett.*** The attractive interaction between ^{3}He atoms turns out to be the spin-spin interaction.

* For John Bardeen (1908–1991), this was his second, after he had got the first Nobel prize in 1956, shared with William Shockley and Walter Brattain, for the discovery of the transistor. The first Nobel prize was a problem for Bardeen, as he needed time to work on superconductivity. In an example to many, he reduced the tam-tam around himself to a minimum, so that he could work as much as possible on the problem of superconductivity. By the way, Bardeen is topped by Frederick Sanger and by Marie Curie. Sanger first won a Nobel prize in chemistry in 1958 by himself and then won a second one shared with Walter Gilbert in 1980; Marie Curie first won one with her husband and a second one by herself, though in two different fields.

** They received the Nobel prize in 1996 for this discovery.

*** Aage Bohr and Ben Mottelson received the Nobel Prize in 1975, Anthony Leggett in 2003.

In superfluids, like in ordinary fluids, one can distinguish between laminar and turbulent flow. The transition between the two regimes is mediated by the behaviour of vortices. But in superfluids, vortices have properties that do not appear in normal fluids. In the superfluid ^{3}He-B phase, vortices are *quantized*: vortices only exist in integer multiples of the
Ref. 799 elementary circulation $h/2m_{^3\mathrm{He}}$. Present research is studying how these vortices behave and how they induce the transition form laminar to turbulent flows.

In recent years, studying the behaviour of *gases* at lowest temperatures has become very popular. When the temperature is so low that the de Broglie wavelength is comparable to the atom-atom distance, bosonic gases form a Bose-Einstein condensate. The first one were realized in 1995 by several groups; the group around Eric Cornell and Carl Wieman used ^{87}Rb, Rand Hulet and his group used ^{7}Li and Wolfgang Ketterle and his group used ^{23}Na. For fermionic gases, the first degenerate gas, ^{40}K, was observed in 1999 by the group around Deborah Jin. In 2004, the same group observed the first gaseous fermi condensate, after the potassium atoms paired up.

Quantized conductivity

Ref. 800 In 1996, the Spanish physicist J.L. Costa–Krämer and his colleagues performed a simple experiment. They put two metal wires on top of each other on a kitchen table and attached a battery, a 10 kΩ resistor and a storage oscilloscope to them. Then they measured the electrical current while knocking on the table. In the last millisecond before the wires detach, the conductivity and thus the electrical current diminished in regular steps of a 7 μA, as can easily be seen on the oscilloscope. This simple experiment could have beaten, if it had been performed a few years earlier, a number of enormously expensive experiments which discovered this quantization at costs of several million euros each, using complex set-ups and extremely low temperatures.

In fact, quantization of conductivity appears in any electrical contact with a small cross section. In such situations the quantum of action implies that the conductivity can only
Challenge 1277 ny be a multiple of $2e^2/\hbar \approx 1/12\,906\,1/\Omega$. Can you confirm this result?

Note that electrical conductivity can be as small as required; only *quantized* electrical conductivity has the minimum value of $2e^2/\hbar$.

The fractional quantum Hall effect

The fractional quantum Hall effect is one of the most intriguing discoveries of materials science. The effect concerns the flow of electrons in a two-dimensional surface. In 1982,
Ref. 802 Robert Laughlin predicted that in this system one should be able to observe objects with
Ref. 801 electrical charge $e/3$. This strange and fascinating prediction was indeed verified in 1997.

The story begins with the discovery by Klaus von Klitzing of the quantum Hall ef-
Ref. 804 fect. In 1980, Klitzing and his collaborators found that in two-dimensional systems at low temperatures – about 1 K – the electrical conductance S is quantized in multiples of the quantum of conductance

$$S = n\,\frac{e^2}{\hbar}\ . \tag{572}$$

The explanation is straightforward: it is the quantum analogue of the classical Hall effect, which describes how conductance varies with applied magnetic field. Von Klitzing received the Nobel prize for physics for the discovery, since the effect was completely unexpected, allows a highly precise measurement of the fine structure constant and also allows one to build detectors for the smallest voltage variations measurable so far.

Two years later, it was found that in extremely strong magnetic fields the conductance could vary in steps *one third* that size. Shortly afterwards, even stranger numerical fractions were also found. In a landmark paper, Robert Laughlin explained all these results by assuming that the electron gas could form collective states showing quasiparticle excitations with a charge $e/3$. This was confirmed 15 years later and earned him a Nobel price as well. We have seen in several occasions that quantization is best discovered through noise measurements; also in this case, the clearest confirmation came from electrical current noise measurements. How can we imagine these excitations?

Ref. 803
Ref. 802

– CS – explanation to be inserted – CS –

What do we learn from this result? Systems in two dimensions have states which follow different rules than systems in three dimensions. Can we infer something about quarks from this result? Quarks are the constituents of protons and neutrons, and have charges $e/3$ and $2e/3$. At this point we need to stand the suspense, as no answer is possible; we come back to this issue later on.

Lasers and other spin-one vector boson launchers

Photons are vector bosons; a lamp is thus a vector boson launcher. All lamps fall into one of three classes. Incandescent lamps use emission from a hot solid, gas discharge lamps use excitation of atoms, ions or molecules through collision, and solid state lamps generate (cold) light through recombination of charges in semiconductors.

Ref. 818

Most solid state lights are light emitting diodes. The large progress in brightness of light emitting diodes could lead to a drastic reduction in future energy consumption, if their cost is lowered sufficiently. Many engineers are working on this task. Since the cost is a good estimate for the energy needed for production, can you estimate which lamp is the most friendly to the environment?

Challenge 1278 n

Nobody thought much about lamps, until Albert Einstein and a few other great physicists came along, such as Theodore Maiman, Hermann Haken and several others that got the Nobel prize with their help. In 1916, Einstein showed that there are two types of sources of light – or of electromagnetic radiation in general – both of which actually 'create' light. He showed that every lamp whose brightness is turned up high enough will change behaviour when a certain intensity threshold is passed. The main mechanism of light emission then changes from spontaneous emission to *stimulated emission*. Nowadays such a special lamp is called a *laser*. (The letters 'se' in laser are an abbreviation of 'stimulated emission'.) After a passionate worldwide research race, in 1960 Maiman was the first to build a laser emitting visible light. (So-called *masers* emitting microwaves were already known for several decades.) In summary, Einstein and the other physicists showed that lasers are lamps which are sufficiently turned up. Lasers consist of some light producing and amplifying material together with a mechanism to pump energy into it. The material can be a gas, a liquid or a solid; the pumping process can use electrical

Table 62 A selection of lamps

Lamp type	Highest brightness (2003)	Lowest cost (2003)	Lifetime (2003)
Incandescent lamps			
tungsten wire light bulbs, halogen lamps	25 lm/W	0.1 cent/lm	700 h
Gas discharge lamps			
oil lamps, candle			
neon lamps			
mercury lamps	100 lm/W	... cent/lm	... h
metal halogenide lamps (ScI_3 or 'xenon', NaI, DyI_3, HoI_3, TmI_5)	110 lm/W	... cent/lm	... h
sodium low pressure lamps	120 lm/W	... cent/lm	... h
sodium high pressure lamps	200 lm/W	... cent/lm	... h
Recombination lamps			
firefly			*c.* 500 h
light emitting diodes	100 lm/W	10 cent/lm	10 000 h
He-Ne laser	550 lm/W	2000 cent/lm	300 h
Ideal white lamp	*c.* 300 lm/W	n.a.	n.a.
Ideal coloured lamp	683 lm/W	n.a.	n.a.

current or light. Usually, the material is put between two mirrors, in order to improve the efficiency of the light production. Common lasers are semiconductor lasers (essentially highly pumped LEDs or light emitting diodes), He–Ne lasers (highly pumped neon lamps), liquid lasers (essentially highly pumped fire flies) and ruby lasers (highly pumped luminescent crystals).

Lasers produce radiation in the range from microwaves and extreme ultraviolet. They have the special property of emitting *coherent* light, usually in a collimated beam. Therefore lasers achieve much higher light intensities than lamps, allowing their use as tools. In modern lasers, the coherence length, i.e. the length over which interference can be observed, can be thousands of kilometres. Such high quality light is used e.g. in gravitational wave detectors.

People have become pretty good at building lasers. Lasers are used to cut metal sheets up to 10 cm thickness, others are used instead of knives in surgery, others increase surface hardness of metals or clean stones from car exhaust pollution. Other lasers drill holes in teeth, measure distances, image biological tissue or grab living cells. Most materials can be used to make lasers, including water, beer and whiskey.

Some materials amplify light so much that end mirrors are not necessary. This is the case for nitrogen lasers, in which nitrogen, or simply air, is used to produce a UV beam. Even a laser made of a single atom (and two mirrors) has been built; in this example,

Ref. 824 only eleven photons on average were moving between the two mirrors. Quite a small lamp. Also lasers emitting light in two dimensions have been built. They produce a light plane instead of a light beam.

Lasers have endless applications. Lasers read out data from compact disks (CDs), are used in the production of silicon integrated circuits, and transport telephone signals; we Page 230 already encountered lasers that work as loudspeakers. Most promising are the applications of femtosecond laser pulses. Femtosecond pulses generate high-temperature plasmas in the materials they propagate, even in air. Such short pulses can be used to cut material without heating it. Recently such lasers have been used to guide lightning along Ref. 819 a predetermined path and seem promising candidates for laser ligtning rods.

Can two photons interfere?

In 1930, Dirac made the famous statement already mentioned above:*

> Each photon interferes only with itself. Interference between two different photons never occurs.

Often this statement is misinterpreted as implying that two separate photon *sources* cannot interfere. It is almost unbelievable how this false interpretation has spread through Ref. 820 the literature. Everybody can check that this statement is incorrect with a radio: two distant radio stations transmitting on the same frequency lead to beats in amplitude, i.e. to *wave interference*. (This should not to be confused with the more common *radio interference*, with usually is simply a superposition of intensities.) Radio transmitters are coherent sources of photons, and any radio receiver shows that two such sources can indeed interfere.

In 1949, interference of two different sources has been demonstrated with microwave beams. Numerous experiments with two lasers and even with two thermal light sources Ref. 820 have shown light interference from the fifties onwards. Most cited is the 1963 experiment by Magyar and Mandel; they used two ruby lasers emitting light pulses and a rapid shutter camera to produce spatial interference fringes.

However, all these experimental results do not contradict the statement by Dirac. Indeed, two photons cannot interfere for several reasons.

- Interference is a result of space-time propagation of waves; photons appear only when the energy–momentum picture is used, mainly when interaction with matter takes place. The description of space-time propagation and the particle picture are mutually exclusive – this is one aspect of the complementary principle. Why does Dirac seem to mix the two in his statement? Dirac employs the term 'photon' in a very general sense, as quantized state of the electromagnetic field. When two coherent beams are superposed, the quantized entities, the photons, cannot be ascribed to either of the sources. Interference results from superposition of two coherent states, not of two particles.
- Interference is only possible if one cannot know where the detected photon comes from. The quantum mechanical description of the field in a situation of interference never

* See the famous, beautiful but difficult textbook P.A.M. Dirac, *The Principles of Quantum Mechanics*, Clarendon Press, Oxford, 1930, page 9.

Figure yet to be included

Figure 324 An electron hologram

allows to ascribe photons of the superposed field to one of the sources. In other words, if you can say from which source a detected photon comes from, you *cannot* observe interference.

▪ Interference between two beams requires a fixed phase between them, i.e. an uncertain particle number; in other words, interference is only possible if the photon number for each of the two beams is unknown.

A better choice of words is to say that interference is always between two (indistinguishable) states, or if one prefers, between two possible (indistinguishable) histories, but never between two particles. In summary, two different electromagnetic beams can interfere, but not two different photons.

Can two electron beams interfere?

Do coherent electron sources exist? Yes, as it is possible to make holograms with electron beams. However, electron coherence is only transversal, not longitudinal. Transversal coherence is given by the possible size of wavefronts with fixed phase. The limit of this size is given by the interactions such a state has with its environment; if the interactions are weak, matter wave packets of several metres of size can be produced, e.g. in particle colliders, where energies are high and interaction with matter is low.

Actually, the term transversal coherence is a fake. The ability to interfere with oneself is not the definition of coherence. Transversal coherence only expresses that the source size is small. Both small lamps (and lasers) can show interference when the beam is split and recombined; this is not a proof of coherence. Similarly, monochromaticity is not a proof for coherence either.

A state is called *coherent* if it possesses a well-defined phase throughout a given domain of space or time. The size of that region or of that time interval defines the degree of coherence. This definition yields coherence lengths of the order of the source size for small 'incoherent' sources. Nevertheless, the size of an interference pattern, or the distance d between its maxima, can be much larger than the coherence length l or the source size s.

In summary, even though an electron can interfere with itself, it cannot interfere with a second one. Uncertain electron numbers are needed to see a macroscopic interference pattern. That is impossible, as electrons (at usual energies) carry a conserved charge.

– CS – sections on transistors and superconductivity to be added – CS –

Figure will be added in the future

Figure 325 Ships in a swell

Challenges and dreams about quantum technology

Many challenges in applied quantum physics remain, as quantum effects seem to promise to realize many age-old technological dreams.

- Challenge 1279 d Is it possible to make A4-size, thin and flexible colour displays for an affordable price?
- Challenge 1280 r Will there ever be desktop laser engravers for 2000 Euro?
- Challenge 1281 r Will there ever be room-temperature superconductivity?
- Challenge 1282 n Will there ever be teleportation of everyday objects?
- Challenge 1283 d Will there ever be applied quantum cryptology?
- Will there ever be printable polymer electronic circuits, instead of lithographically Challenge 1284 d patterned silicon electronics as is common now?
- Challenge 1285 r Will there ever be radio-controlled flying toys in the size of insects?

27. Quantum electrodynamics – the origin of virtual reality

The central concept the quantum theory introduces in the description of nature is the idea of *virtual particles*. Virtual particles are short-lived particles; they owe their existence exclusively to the quantum of action. Because of the quantum of action, they do not need to follow the energy-mass relation that special relativity requires of normal, real particles. Virtual particles can move faster than light and can move backward in time. Despite these strange properties, they have many observable effects.

Ships, mirrors and the Casimir effect

When two parallel ships roll in a big swell, *without* even the slightest wind blowing, they will attract each other. This effect was well known up to the nineteenth century, when many places still lacked harbours. Shipping manuals advised captains to let the ships be Ref. 808 pulled apart using a well-manned rowing boat.

Waves induce oscillations of ships because a ship absorbs energy from the waves. When oscillating, the ship also emits waves. This happens mainly towards the two sides of the ship. As a result, for a single ship, the wave emission has no net effect on its position. Now imagine that two parallel ships oscillate in a long swell, with a wavelength much larger than the distance between the ships. Due to the long wavelength, the two ships will oscillate in phase. The ships will thus not be able to absorb energy from each other. As a result, the energy they radiate towards the outside will push them towards each other.

The effect is not difficult to calculate. The energy of a rolling ship is

$$E = mgh\,\alpha^2/2 \tag{573}$$

where α is the roll angle amplitude, m the mass of the ship and $g = 9,8\,\mathrm{m/s^2}$ the acceleration due to gravity. The *metacentric height* h is the main parameter characterizing a ship, especially a sailing ship; it tells with what torque the ship returns to the vertical when inclined by an angle α. Typically, one has h =1.5 m.

When a ship is inclined, it will return to the vertical by a damped oscillation. A damped oscillation is characterized by a period T and a quality factor Q. The *quality factor* is the number of oscillations the system takes to reduce its amplitude by a factor $e = 2.718$. If the quality factor Q of an oscillating ship and its oscillation period T are given, the radiated power W is

$$W = 2\pi\frac{E}{QT}\ . \tag{574}$$

We saw above that radiation pressure is W/c, where c is the wave propagation velocity. For water waves, we have the famous relation

$$c = \frac{gT}{2\pi}\ . \tag{575}$$

Assuming that for two nearby ships each one completely absorbs the power emitted from the other, we find that the two ships are attracted towards each other following

$$ma = m2\pi^2\frac{h\alpha^2}{QT^2}\ . \tag{576}$$

Inserting typical values such as $Q = 2.5$, T =10 s, α =0.14 rad and a ship mass of 700 tons, we get about 1.9 kN. Long swells thus make ships attract each other. The intensity of the attraction is comparatively small and can indeed be overcome with a rowing boat. On the other hand, even the slightest wind will damp the oscillation amplitude and have other effects that will avoid the observation of the attraction.

Sound waves or noise in air can have the same effect. It is sufficient to suspend two metal plates in air and surround them by loudspeakers. The sound will induce attraction (or repulsion) of the plates, depending on whether the sound wavelength cannot (or can) be taken up by the other plate. Ref. 809

In 1948, the Dutch physicist Hendrik Casimir made one of the most spectacular predictions of quantum theory: he predicted a similar effect for metal plates in vacuum. Casimir, who worked at the Dutch Electronics company Philips, wanted to understand why it was so difficult to build television tubes. Television screens are made by deposing small neutral particles on glass, but Casimir observed that the particles somehow attracted each other. Casimir got interested in understanding how neutral particles interact. During these theoretical studies he discovered that two neutral mirrors (or metal plates) would attract each other even in complete vacuum. This is the famous *Casimir effect*. Casimir also determined the attraction strength between a sphere and a plate, and between two spheres. In fact, all conducting bodies attract each other in vacuum, with a force depending on their geometry.

In all these situations, the role of the sea is taken by the zero-point fluctuations of the electromagnetic field, the role of the ships by the mirrors. Casimir understood that the space between two parallel mirrors, due to the geometrical constraints, had different zero-point fluctuations that the free vacuum. Like two ships, the result would be the attraction of the mirrors.

Casimir predicted that the attraction for two mirrors of mass m and surface A is given by

$$\frac{ma}{A} = \frac{\pi^3}{120} \frac{\hbar c}{d^4} . \qquad (577)$$

The effect is a pure quantum effect; in classical electrodynamics, two neutral bodies do not attract. The effect is small; it takes some dexterity to detect it. The first experimental check was by Marcus Sparnaay, Casimir's colleague at Philips, in 1958. Two beautiful high-precision measurements of the Casimir effect were performed in 1997 by Lamoreaux and in 1998 by Mohideen and Roy; they confirmed Casimir's prediction with a precision of 5 % and 1 % respectively.

Ref. 810
Ref. 811

In a cavity, spontaneous emission is suppressed, if it is smaller than the wavelength of the emitted light! This effect has also been observed. It confirms the old saying that spontaneous emission is emission stimulated by the zero point fluctuations.

The Casimir effect thus confirms the existence of the zero-point fluctuations of the electromagnetic field. It confirms that quantum theory is valid also for electromagnetism.

The Casimir effect between two spheres is proportional to $1/r^7$ and thus is much weaker than between two parallel plates. Despite this strange dependence, the fascination of the Casimir effect led many amateur scientists to speculate that a mechanism similar to the Casimir effect might explain gravitational attraction. Can you give at least three arguments why this is impossible, even if the effect had the correct distance dependence?

Challenge 1286 n

Like the case of sound, the Casimir effect can also produce repulsion instead of attraction. It is sufficient that one of the two materials be perfectly permeable, the other a perfect conductor. Such combinations repel each other, as Timothy Boyer discovered in 1974.

Ref. 812
Ref. 813

The Casimir effect bears another surprise: between two metal plates, the speed of light changes and can be larger than c. Can you imagine what exactly is meant by 'speed of light' in this context?

Challenge 1287 n

The Banach–Tarski paradox for vacuum

It implies that there is a specific energy density that can be described to the vacuum. This seems obvious. However, the statement has a dramatic consequence: space-time cannot be continuous!

The reasoning is simple. If the vacuum were continuous, we could make use of the Banach–Tarski paradox and split, without any problem, a ball of vacuum into two balls

Page 54

of vacuum, each with the same volume. In other words, one ball with energy E could not be distinguished from two balls of energy $2E$. This is impossible.

The Gedanken experiment tells us something important. In the same way that we used the argument to show that chocolate (and any other matter) cannot be continuous, we can now deduce that the vacuum cannot be either. However, we have no details yet. In the same way that matter turned out to possess an intrinsic scale, we can guess that this happens also to the vacuum. Vacuum has an intrinsic scale; it is not continuous. We will have to wait for the third part of the text to find out more. There, the structure of the vacuum will turn out to be even more interesting than that of matter.

Page 202

The Lamb shift

In the 1947, the measurements of the spectrum of hydrogen had yielded another effect due to virtual particles. Willis Lamb (1913–) found that the $2S_{1/2}$ energy level in atomic hydrogen lies slightly above the $2P_{1/2}$ level. This is in contrast to the calculation performed above, where the two levels are predicted to have the same energy. In reality, they have an energy difference of 1057.864 MHz or 4.3 μeV. This discovery had important consequences for the description of quantum theory and yielded Lamb a share of the 1955 Nobel Prize.

Page 706

The reason for the difference is an unnoticed approximation performed in the simple solution above. There are two equivalent ways to explain it. One is to say that the calculation neglects the coupling terms between the Dirac equation and the Maxwell equations. This explanation lead to the first calculations of the Lamb shift, around the year 1950. The other explanation is to say that the calculation neglects virtual particles. In particular, the calculation neglects the virtual photons emitted and absorbed during the motion of the electron around the nucleus. This is the explanation in line with the general vocabulary of quantum electrodynamics. QED is perturbative approach to solve the coupled Dirac and Maxwell equations.

The QED Lagrangian

– CS – section on the QED Lagrangian to be added – CS –

Interactions and virtual particles

The electromagnetic interaction is exchange of virtual photons. So how can the interaction be attractive? At first sight, any exchange of virtual photons should drive the electrons from each other. However, this is not correct. The momentum of virtual photons does not have to be in the direction of its energy flow; it can also be in opposite direction.* Obviously, this is only possible within the limits provided by the indeterminacy principle.

* One of the most beautiful booklets on quantum electrodynamics which makes this point remains the text by RICHARD FEYNMAN, *QED: the Strange Theory of Light and Matter*, Penguin Books, 1990.

Vacuum energy

The strangest result of quantum field theory is the energy density of the vacuum.

– CS – More to be written – CS –

Moving mirrors

Mirrors also work when in motion; in contrast, walls that produce echoes do not work at all speeds. Walls do not produce echoes if one moves faster than sound. However, mirrors always produce an image. This observation shows that the speed of light is the same for any observer. Can you detail the argument? Challenge 1288 n

Mirrors also differ from tennis rackets. We saw that mirrors cannot be used to change the speed of the light they hit, in contrast to what tennis rackets can do with balls. This observation shows that the speed of light is also a limit velocity. In short, the simple existence of mirrors is sufficient to derive special relativity.

But there are more interesting things to be learned from mirrors. We only have to ask whether mirrors work when they undergo accelerated motion. This issue yields a surprising result.

In the 1970s, quite a number of researchers found that there is no vacuum for accelerated observers. This effect is called *Filling–Davies–Unruh effect* or sometimes the *dynamical Casimir effect*. As a consequence, a mirror in accelerated motion reflects the fluctuations it encounters and reflects them. In short, *an accelerated mirror emits light*. Unfortunately, the intensity is so weak that it has not been measured up to now. We will explore the issue in more detail below. Can you explain why accelerated mirrors emit light, but not matter? Page 812 Challenge 1289 n

Photon hitting photons

When virtual particles are taken into account, light beams can 'bang' onto each other. This result is in contrast to classical electrodynamics. Indeed, QED shows that the virtual electron-positron pairs allow photons to hit each other. And such pairs are found in any light beam.

However, the cross section is small. When two beams cross, most photons will pass undisturbed. The cross section A is approximately

$$A \approx \frac{973}{10\,125\pi}\alpha^4\Big(\frac{\hbar}{m_e c}\Big)^2\Big(\frac{\hbar\omega}{m_e c^2}\Big)^6 \qquad (578)$$

for the case that the energy $\hbar\omega$ of the photon is much smaller than the rest energy $m_e c^2$ of the electron. This value is about 18 orders of magnitude smaller than what was measurable in 1999; the future will show whether the effect can be observed for visible light. However, for high energy photons these effects are observed daily in particle accelerators. In these cases one observes not only interaction through virtual electron–antielectron pairs, but also through virtual muon–antimuon pairs, virtual quark–antiquark pairs, and much more.

Everybody who consumes science fiction knows that matter and antimatter annihilate and transform into pure light. In more detail, a matter particle and an antimatter particle annihilate into two or more photons. More interestingly, quantum theory predicts that the opposite process is also possible: photons hitting photons can produce matter!

Ref. 814 In 1997, this was also confirmed experimentally. At the Stanford particle accelerator, photons from a high energy laser pulse were bounced off very fast electrons. In this way, the reflected photons acquired a large energy, when seen in the inertial frame of the experimenter. The original pulse, of 527 nm or 2.4 eV green light, had a peak power density of 10^{22} W/m^2, about the highest achievable so far. That is a photon density of 10^{34} /m^3 and an electric field of 10^{12} V/m, both of which were record values at the time. When this laser pulse was reflected off a 46.6 GeV electron beam, the returning photons had an energy of 29.2 GeV and thus had become intense gamma rays. These gamma rays then collided with still incoming green photons and produced electron–positron pairs by the reaction

Challenge 1290 ny

$$\gamma_{29.2} + n\,\gamma_{\text{green}} \to e^{+} + e^{-} \tag{579}$$

for which both final particles were detected by special apparatuses. The experiment thus showed that light can hit light in nature, and above all, that doing so can produce matter. This is the nearest one can get to the science fiction idea of light swords or of laser swords banging onto each other.

Is the vacuum a bath?

If the vacuum is a sea of virtual photons and particle–antiparticle pairs, vacuum could be suspected to act as a bath. In general, the answer is negative. Quantum field theory works because the vacuum is *not* a bath for single particles. However, there is always an exception. For dissipative systems made of many particles, such as electrical conductors, the vacuum *can* act as a viscous fluid. Irregularly shaped, neutral, but conducting bodies can emit photons when accelerated, thus damping such type of motion. This is due to the Fulling–Davies–Unruh effect, also called the *dynamical Casimir effect*, as described above. The damping depends on the shape and thus also on the direction of the body's motion.

Ref. 815

Page 221 In 1998, Gour and Sriramkumar even predicted that Brownian motion should also appear for an imperfect, i.e. partly absorbing mirror placed in vacuum. The fluctuations of the vacuum should produce a mean square displacement

Ref. 816

$$\langle d^2 \rangle = \hbar / mt \tag{580}$$

increasing linearly with time; however, the extremely small displacements produced this way seem out of experimental reach so far. But the result is not a surprise. Are you able to give another, less complicated explanation for it?

Challenge 1291 ny

Renormalization – why is an electron so light?

– CS – section on renormalization to be added – CS –

Sometimes it is claimed that the infinities appearing in quantum electrodynamics in the intermediate steps of the calculation show that the theory is incomplete or wrong. However, this type of statement would imply that classical physics is also incomplete or wrong, on the ground that in the definition of the velocity v with space x and time t, namely

$$v = \frac{\mathrm{d}x}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \lim_{\Delta t \to 0} \Delta x \frac{1}{\Delta t} , \qquad (581)$$

one gets an infinity as intermediate step. Indeed, $\mathrm{d}t$ being vanishingly small, one could argue that one is dividing by zero. Both arguments show the difficulty to accept that the result of a limit process can be a finite quantity even if infinite quantities appear in it. The parallel with the definition of the velocity is closer than it seems; both 'infinities' stem from the assumption that space-time is continuous, i.e. infinitely divisible. The infinities necessary in limit processes for the definition of differentiation, of integration or for the renormalization scheme appear only when space-time is approximated as a complete set, or as physicists say, as a 'continuous' set.

On the other hand, the conviction that the appearance of an infinity might be a sign of incompleteness of a theory was an interesting development in physics. It shows how uncomfortable many physicists had become with the use of infinity in our description of nature. Notably, this was the case for Dirac himself, who, after having laid in his youth the basis of quantum electrodynamics, has tried for the rest of his life to find a way, without success, to change the theory so that infinities are avoided.* Ref. 817

Renormalization is a procedure that follows from the requirement that continuous space-time and gauge theories must work together. In particular, it follows form the requirement that the particle concept is consistent, i.e. that perturbation expansions are possible.

Curiosities and fun challenges of quantum electrodynamics

Motion is an interesting topic, and when a curious person asks a question about it, most of the time quantum electrodynamics is needed for the answer. Together with gravity, quantum electrodynamics explains almost all of our everyday experience, including numerous surprises. Let us have a look at some of them.

- There is a famous riddle asking how far the last card (or the last brick) of a stack can hang over the edge of a table. Of course, only gravity, no glue or any other means is allowed to keep the cards on the table. After you solved the riddle, can you give the solution in case that the quantum of action is taken into account? Challenge 1292 n

- Quantum electrodynamics explains why there are only a *finite* number of different atom types. In fact, it takes only two lines to prove that pair production of electron–antielectron pairs make it impossible that a nucleus has more than about 137 protons. Can you show this? The effect at the basis of this limit, the polarization of the vacuum, also plays a role in much larger systems, such as charged black holes, as we will see shortly. Ref. 821 Challenge 1293 n Page 815

* Not long after his death, his wish has been fulfilled, although in a different manner that he envisaged. The third part of this mountain ascent will show the way out of the issue.

▪ Taking 91 of the 92 electrons off an uranium atom allows researchers to check whether the innermost electron still is described by QED. The electric field near the uranium nucleus, 1 EV/m is near the threshold for spontaneous pair production. The field is the highest constant field producible in the laboratory, and an ideal testing ground for precision QED experiments. The effect of virtual photons is to produce a Lamb shift; even in these extremely high fields, the value fits with the predictions.

Ref. 822

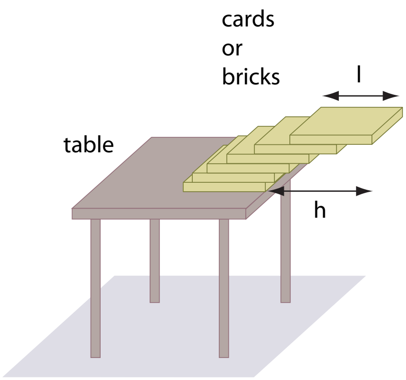


Figure 326 What is the maximum possible value of h/l?

▪ Is there a critical magnetic field in nature, like there is a critical electric field, limited by spontaneous pair production?

Challenge 1294 ny

▪ In classical physics, the field energy of a point-like charged particle, and hence its mass, was predicted to be infinite. QED effectively *smears out* the charge of the electron over its Compton wavelength, so that in the end the field energy contributes only a small correction to its total mass. Can you confirm this?

Page 554
Challenge 1295 ny

▪ Microscopic evolution can be pretty slow. Light, especially when emitted by single atoms, is always emitted by some metastable state. Usually, the decay times, being induced by the vacuum fluctuations, are much shorter than a microsecond. However, there are metastable atomic states with a lifetime of ten years: for example, an ytterbium ion in the $^2F_{7/2}$ state achieves this value, because the emission of light requires an octupole transition, in which the angular momentum changes by $3\hbar$; this is an extremely unlikely process.

Ref. 823

▪ Microscopic evolution can be pretty fast. Can you imagine how to deduce or to measure the speed of electrons inside atoms? And inside metals?

Challenge 1296 n

▪ Take a horseshoe. The distance between the two ends is not fixed, since otherwise their position and velocity would be known at the same time, contradicting the indeterminacy relation. Of course, this reasoning is also valid for any other solid object. In short, both quantum mechanics and special relativity show that rigid bodies do not exist, albeit for different reasons.

▪ Have you ever admired a quartz crystal or some other crystalline material? The beautiful shape and atomic arrangement has formed spontaneously, as a result of the motion of atoms under high temperature and pressure, during the time that the material was deep under the Earth's surface. The details of crystal formation are complex and interesting.

For example, are regular crystal lattices *energetically* optimal? This simple question leads to a wealth of problems. We might start with the much simpler question whether a regular dense packing of spheres is the most *dense* possible. Its density is $\pi/\sqrt{18}$, i.e. a bit over 74 %. Even though this was conjectured to be the maximum possible value already in 1609 by Johannes Kepler, the statement was proven only in 1998 by Tom Hales. The proof is difficult because in small volumes it is possible to pack spheres up to almost 78 %. To show that over large volumes the lower value is correct is a tricky business.

Challenge 1297 n
Ref. 825

Next, does a regular crystal of solid spheres, in which the spheres do *not* touch, have the lowest possible entropy? This simple problem has been the subject of research only in the 1990s. Interestingly, *for low temperatures*, regular sphere arrangements indeed show

Figure 327 Some snow flakes (© Furukawa Yoshinori)

the largest possible entropy. At low temperatures, spheres in a crystal can oscillate around their average position and be thus more disordered than if they were in a liquid; in the liquid state the spheres would block each other's motion and would not allow to show disorder at all.

This many similar results deduced from the research into these so-called *entropic forces* show that the transition from solid to liquid is – at least in part – simply a geometrical effect. For the same reason, one gets the surprising result that even slightly repulsing spheres (or atoms) can form crystals and melt at higher temperatures. These are beautiful examples of how classical thinking can explain certain material properties, using from quantum theory only the particle model of matter.

Ref. 826

But the energetic side of crystal formation provides other interesting questions. Quantum theory shows that it is possible that *two* atoms repel each other, while *three* attract each other. This beautiful effect was discovered and explained by Hans–Werner Fink in 1984. He studied rhenium atoms on tungsten surfaces and showed, as observed, that they cannot form dimers – two atoms moving closeby – but readily form trimers. This is an example contradicting classical physics; the effect is impossible if one pictures atoms as immutable spheres, but becomes possible when one remembers that the electron clouds around the atoms rearrange depending on their environment.

Ref. 827

For an exact study of crystal energy, the interactions between all atoms have to be included. The simplest question is to determine whether a regular array of alternatively charged spheres has lower energy than some irregular collection. Already such simple questions are still topic of research; the answer is still open.

Another question is the mechanism of face formation in crystals. Can you confirm that crystal faces are those planes with the *slowest* growth speed, because all fast growing planes are eliminated? The finer details of the process form a complete research field in itself.

Challenge 1298 n
Ref. 828

However, not always the slowest growing planes win out. Figure 327 shows some well-known exceptions. Explaining such shapes is possible today, and Furukawa Yoshinori is one of the experts in the field, heading a dedicated research team. Indeed, there remains the question of symmetry: why are crystals often symmetric, such as snowflakes, instead of asymmetric? This issue is a topic of self-organization, as mentioned already in the section of classical physics. It turns out that the symmetry is an automatic result of the way molecular systems grow under the combined influence of diffusion and nonlinear processes. The details are still a topic of research.

Ref. 829
Page 232

▪ A similar breadth of physical and mathematical problems are encountered in the

Ref. 830 study of liquids and polymers. The ordering of polymer chains, the bubbling of hot water, the motion of heated liquids and the whirls in liquid jets show complex behaviour that can be explained with simple models. Turbulence and self-organization will be a fascinating research field for many years to come.

▪ The ways people handle single atoms with electromagnetic fields is a beautiful example of modern applied technologies. Nowadays it is possible to levitate, to trap, to excite, to photograph, to deexcite and to move single atoms just by shining light onto them. In 1997, the Nobel prize in physics has been awarded to the originators of the field.

▪ In 1997, a Czech group built a quantum version of the Foucault pendulum, using the superfluidity of helium. (Ref. 831) In this beautiful piece of research, they cooled a small ring of fluid helium below the temperature of 0.28 K, below which the helium moves without friction. In such situations it thus can behave like a Foucault pendulum. With a clever arrangement, it was possible to measure the rotation of the helium in the ring using phonon signals, and to show the rotation of the Earth.

▪ If an electrical wire is sufficiently narrow, its electrical conductance is quantized in steps of $2e^2/\hbar$. The wider the wire, the more such steps are added to its conductance. Can you explain the effect? (Challenge 1299 n) By the way, quantized conductance has also been observed for light and for phonons. (Ref. 832)

▪ An example of modern research is the study of hollow atoms, i.e. atoms missing a number of inner electrons. They have been discovered in 1990 by J.P. Briand and his group. They appear when a completely ionized atom, i.e. one without any electrons, is brought in contact with a metal. The acquired electrons then orbit on the outside, leaving the inner shells empty, in stark contrast with usual atoms. Such hollow atoms can also be formed by intense laser irradiation. (Ref. 833)

▪ In the past, the description of motion with formulae was taken rather seriously. Before computers appeared, only those examples of motion were studied which could be described with simple formulae. It turned out that Galilean mechanics cannot solve the three-body problem, special relativity cannot solve the two-body problem, general relativity the one-body problem and quantum field theory the zero-body problem. It took some time to the community of physicists to appreciate that understanding motion does not depend on the description by formulae, but on the description by clear equations based on space and time.

Challenge 1300 n ▪ Can you explain why mud is not clear?

▪ Photons not travelling parallel to each other attract each other through gravitation and thus deflect each other. Could two such photons form a bound state, a sort of atom of light, in which they would circle each other, provided there were enough empty space for this to happen? (Challenge 1301 n)

Challenge 1302 n ▪ Can the universe ever have been smaller than its own Compton wavelength?

In fact, quantum electrodynamics, or QED, provides a vast number of curiosities and every year there is at least one interesting new discovery. We now conclude the theme with a more general approach.

How can one move on perfect ice? – The ultimate physics test

In our quest, we have encountered motion of many sorts. Therefore, the following test – not to be taken too seriously – is the ultimate physics test, allowing to check your understanding and to compare it with that of others.

Imagine that you are on a perfectly frictionless surface and that you want to move to its border. How many methods can you find to achieve this? Any method, so tiny its effect may be, is allowed.

Classical physics provided quite a number of methods. We saw that for rotating ourselves, we just need to turn our arm above the head. For translation motion, throwing a shoe or inhaling vertically and exhaling horizontally are the simplest possibilities. Can you list at least six additional methods, maybe some making use of the location of the surface on Earth? What would you do in space? Challenge 1303 n

Electrodynamics and thermodynamics taught us that in vacuum, heating one side of the body more than the other will work as motor; the imbalance of heat radiation will push you, albeit rather slowly. Are you able to find at least four other methods from these two domains? Challenge 1304 n

General relativity showed that turning one arm will emit gravitational radiation unsymmetrically, leading to motion as well. Can you find at least two better methods? Challenge 1305 n

Quantum theory offers a wealth of methods. Of course, quantum mechanics shows that we actually are always moving, since the indeterminacy relation makes rest an impossibility. However, the average motion can be zero even if the spread increases with time. Are you able to find at least four methods of moving on perfect ice due to quantum effects? Challenge 1306 n

Materials science, geophysics, atmospheric physics and astrophysics also provide ways to move, such as cosmic rays or solar neutrinos. Can you find four additional methods? Challenge 1307 n

Self-organization, chaos theory and biophysics also provide ways to move, when the inner workings of the human body are taken into account. Can you find at least two methods? Challenge 1308 n

Assuming that you read already the section following the present one, on the effects of semiclassical *quantum gravity*, here is an additional puzzle: is it possible to move by accelerating a pocket mirror, using the emitted Unruh radiation? Can you find at least two other methods to move yourself using quantum gravity effects? Can you find one from string theory? Challenge 1309 n

If you want points for the test, the marking is simple. For students, every working method gives one point. Eight points is ok, twelve points is good, sixteen points is very good, and twenty points or more is excellent.* For graduated physicists, the point is given only when a back-of-the-envelope estimate for the ensuing momentum or acceleration is provided.

Summary of quantum electrodynamics

The shortest possible summary of quantum electrodynamics is the following: *matter is made of charged particles which interact through photon exchange in the way described by Figure 328.*

No additional information is necessary. In a bit more detail, quantum electrodynamics starts with elementary *particles*, characterized by their mass, their spin and their charge, and with the *vacuum*, essentially a sea of virtual particle–antiparticle pairs. *Interactions*

* The author keeps track of all answers on the http://www.motionmountain.org web site.

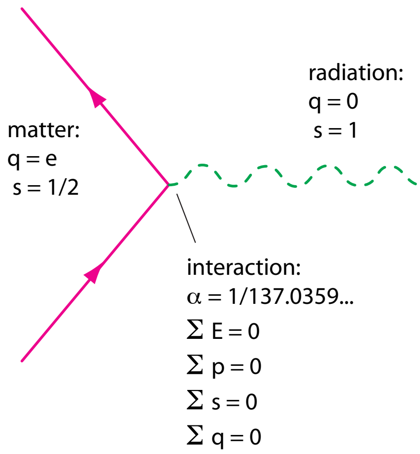


Figure 328 QED as perturbation theory in space-time

between charged particles are described as the exchange of virtual photons, and *decay* is described as the interaction with the virtual photons of the vacuum.

All physical results of QED can be calculated by using the single diagram of Figure 328. As QED is a perturbative theory, the diagram directly describes the first order effects and its composites describe effects of higher order. QED is a perturbative theory.

QED describes all everyday properties of matter and radiation. It describes the *divisibility* down to the smallest constituents, the *isolability* from the environment and the *impenetrability* of matter. It also describes the *penetrability* of radiation. All these properties are due to electromagnetic interactions of constituents and follow from Figure 328. Matter is divisible because the interactions are of finite strength, matter is separable because the interactions are of finite range, and matter is impenetrable because interactions among the constituents increase in intensity when they approach each other, in particular because matter constituents are fermions. Radiation is divisible into photons, and is penetrable because photons are bosons and first order photon-photon interactions do not exist.

Both matter and radiation are made of elementary constituents. These elementary constituents, whether bosons or fermions, are indivisible, isolable, indistinguishable, and point-like.

To describe observations, it is necessary to use quantum electrodynamics in all those situations for which the characteristic dimensions d are of the order of the Compton wavelength

$$d \approx \lambda_{\mathrm{C}} = \frac{h}{m\,c} \,. \qquad (582)$$

In situations where the dimensions are of the order of the de Broglie wavelength, or equivalently, where the action is of the order of the Planck value, simple quantum mechanics is sufficient:

$$d \approx \lambda_{\mathrm{dB}} = \frac{h}{m\,v} \,. \qquad (583)$$

For larger dimensions, classical physics will do.

Together with gravity, quantum electrodynamics explains almost all observations of motion on Earth; QED unifies the description of matter and radiation in daily life. All objects and all images are described by it, including their properties, their shape, their transformations and their other changes. This includes self-organization and chemical or biological. In other words, QED gives us full grasp of the effects and the variety of motion due to electromagnetism.

Open questions in QED

Even though QED describes motion without any discrepancy from experiment, that does not mean that we understand every detail of every example of electric motion. For example, nobody has described the motion of an animal with QED yet.* In fact, there is beautiful and fascinating work going on in many branches of electromagnetism.

Atmospheric physics still provides many puzzles and regularly delivers new, previously unknown phenomena. For example, the detailed mechanisms at the origin of aurorae are still controversial; and the recent unexplained discoveries of discharges *above* clouds should not make one forget that even the precise mechanism of charge separation *inside* clouds, which leads to lightning, is not completely clarified. In fact, all examples of electrification, such as the charging of amber through rubbing, the experiment which gave electricity its name, are still poorly understood. Ref. 835 Ref. 836

Materials science in all its breadth, including the study of solids, fluids, and plasmas, as well as biology and medicine, still provides many topics of research. In particular, the twenty-first century will undoubtedly be the century of the life sciences.

The study of the interaction of atoms with intense light is an example of present research in atomic physics. Strong lasers can strip atoms of many of their electrons; for such phenomena, there are not yet precise descriptions, since they do not comply to the weak field approximations usually assumed in physical experiments. In strong fields, new effects take place, such as the so-called Coulomb explosion. Ref. 838

But also the skies have their mysteries. In the topic of cosmic rays, it is still not clear how rays with energies of 10^{22} eV are produced outside the galaxy. Researchers are intensely trying to locate the electromagnetic fields necessary for their acceleration and to understand their origin and mechanisms. Ref. 837

In the theory of quantum electrodynamics, discoveries are expected by all those who study it in sufficient detail. For example, Dirk Kreimer has found that higher order interaction diagrams built using the fundamental diagram of Figure 328 contain relations to the theory of knots. This research topic will provide even more interesting results in the near future. Ref. 839

Relations to knot theory appear because QED is a *perturbative* description, with the vast richness of its nonperturbative effects still hidden. Studies of QED at high energies, where perturbation is *not* a good approximation and where particle numbers are not conserved, promise a wealth of new insights. We will return to the topic later on.

High energies provide many more questions. So far, the description of motion was based on the idea that measurable quantities can be multiplied and added. This always

* On the other hand, there is beautiful work going on how humans move their limbs; it seems that humans move by combining a small set of fundamental motions. Ref. 834

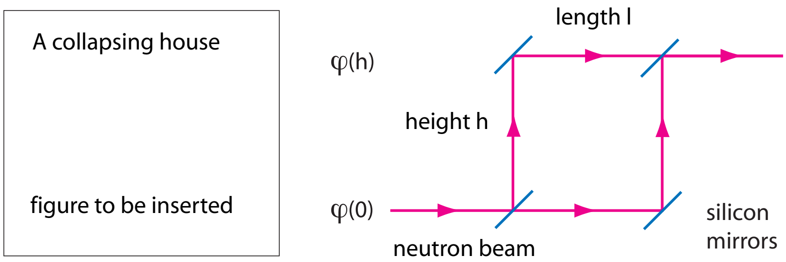


Figure 329 The weakness of gravitation

happens at one space-time point. In mathematical jargon, observables form a *local* algebra. Thus the structure of an algebra contains, implies and follows from the idea that local properties *lead* to local properties. We will discover later on that this basic assumption is wrong at high energies.

We defined special relativity using $v \leqslant c$, general relativity using $L/M \geqslant 4G/c^2$ and quantum theory using $S \geqslant \hbar/2$. How can we define electromagnetism in one statement? This is not known yet. Challenge 1310 r

Many other open issues of more practical nature have not been mentioned. Indeed, by far the largest numbers of physicists get paid for some form of applied QED. However, our quest is the description of the *fundamentals* of motion. So far, we have not achieved it. For example, we still need to understand motion in the realm of atomic nuclei. But before we do that, we take a first glimpse of the strange issues appearing when gravity and quantum theory meet.

28. Quantum mechanics with gravitation – the first approach

Gravitation is a weak effect. Every seaman knows it: storms are the worst part of his life, not gravity. Nevertheless, including gravity into quantum mechanics yields a list of important issues.

In the chapter on general relativity we already mentioned that light frequency changes with height. But for matter wavefunctions, gravity also changes their *phase*. Can you imagine why? The effect was first confirmed in 1975 with the help of neutron interferometers, where neutron beams are brought to interference after having climbed some height h at two different locations. The experiment is shown schematically in Figure 329; it fully confirmed the predicted phase difference Challenge 1311 ny

$$\delta\varphi = \frac{mghl}{\hbar v} \qquad (584)$$

where l is the distance of the two climbs and v and m are the speed and mass of the neutrons. These beautifully simple experiments have confirmed the formula within experimental errors.* Ref. 840

* Due to the influence of gravity on phases of wavefunctions, some people who do not believe in bath induced decoherence have even studied the influence of gravity on the decoherence process of usual quantum

Ref. 842 In the 1990s, similar experiments have even been performed with complete atoms. These set-ups allow to build interferometers so sensitive that local gravity g can be measured with a precision of more than eight significant digits.

Corrections to the Schrödinger equation

In 2002, the first observation of actual quantum states due to gravitational energy was performed. Any particle above the floor should feel the effect of gravity. Ref. 843

In a few words, one can say that because the experimenters managed to slow down neutrons to the incredibly small value of 8 m/s, using grazing incidence on a flat plate they could observe how neutrons climbed and fell back due to gravity with speeds below a few cm/s.

Obviously, the quantum description is a bit more involved. The lowest energy level for neutrons due to gravity is $2.3 \cdot 10^{-31}$ J, or 1.4 peV. To get an impression of it smallness, we can compare it to the value of $2.2 \cdot 10^{-18}$ J or 13.6 eV for the lowest state in the hydrogen atom.

A rephrased large number hypothesis

Despite its weakness, gravitation provides many puzzles. Most famous are a number of curious coincidences that can be found when quantum mechanics and gravitation are combined. They are usually called 'large number hypotheses' because they usually involve large dimensionless numbers. A pretty, but less well known version connects the Planck length, the cosmic horizon, and the number of baryons: Ref. 865

$$(N_{\mathrm{b}})^3 \approx \left(\frac{R_0}{l_{\mathrm{Pl}}}\right)^4 = \left(\frac{t_0}{t_{\mathrm{Pl}}}\right)^4 \approx 10^{244} \tag{585}$$

in which $N_{\mathrm{b}} = 10^{81}$ and $t_0 = 1.2 \cdot 10^{10}$ a were used. There is no known reason why the number of baryons and the horizon size R_0 should be related in this way. This coincidence is equivalent to the one originally stated by Dirac,* namely

$$m_{\mathrm{p}}^3 \approx \frac{\hbar^2}{G c t_0} \ . \tag{587}$$

where m_{p} is the proton mass. This approximate equality seems to suggest that certain microscopic properties, namely the mass of the proton, is connected to some general

Ref. 841 systems in flat space-time. Predictably, the calculated results do not reproduce experiments.

Ref. 866 * The equivalence can be deduced using $G n_{\mathrm{b}} m_{\mathrm{p}} = 1/t_0^2$, which, as Weinberg explains, is required by several cosmological models. Indeed, this can be rewritten simply as Page 964

$$m_0^2/R_0^2 \approx m_{\mathrm{Pl}}^2/R_{\mathrm{Pl}}^2 = c^4/G^2 \ . \tag{586}$$

Together with the definition of the baryon density $n_{\mathrm{b}} = N_{\mathrm{b}}/R_0^3$ one gets Dirac's large number hypothesis, substituting protons for pions. Note that the Planck time and length are defined as $\sqrt{\hbar G/c^5}$ and $\sqrt{\hbar G/c^3}$ and are the natural units of length and time. We will study them in detail in the third part of the mountain ascent.

properties of the universe as a whole. This has lead to numerous speculations, especially since the time dependence of the two sides differs. Some people even speculate whether relations (585) or (587) express some long-sought relation between local and global topological properties of nature. Up to this day,the only correct statement seems to be that they are *coincidences* connected to the time at which we happen to live. But gravity also leads to other quantum surprises.

Ref. 867

Is quantum gravity necessary?

One might think that gravity does not require a quantum description. We remember that we stumbled onto quantum effects because classical electrodynamics implies, in stark contrast with reality, that atoms decay in about 0.1 ns. Classically, an orbiting electron would emit radiation until it falls into the nucleus. Quantum theory is necessary to save the situation.

When the same calculation is performed for the emission of gravitational radiation by orbiting electrons, one finds a decay time of around 10^{37} s. (True?) This extremely large value, trillions of times longer than the age of the universe, is a result of the low emission of gravitational radiation by rotating masses. Therefore, the existence of atoms does not require a quantum theory of gravity.

Challenge 1312 ny
Page 372

Indeed, quantum gravity is unnecessary in every single domain of everyday life. However, quantum gravity is necessary in domains which are more remote, but also more fascinating.

Limits to disorder

> Die Energie der Welt ist constant.
> Die Entropie der Welt strebt einem Maximum zu.*
> Rudolph Clausius

We have already encountered the famous statement by Clausius, the father of the term 'entropy'. Strangely, for over hundred years nobody asked whether there actually exists a theoretical maximum for entropy. This changed in 1973, when Jakob Bekenstein found the answer while investigating the consequences gravity has for quantum physics. He found that the entropy of an object of energy E and size L is bound by

Page 436
Ref. 844

$$S \leqslant EL\frac{k\pi}{\hbar c} \tag{588}$$

for all physical systems. In particular, he deduced that (nonrotating) black holes saturate the bound, with an entropy given by

Challenge 1313 n

$$S = \frac{kc^3}{G\hbar}\frac{A}{4} = \frac{kG}{\hbar c}4\pi M^2 \tag{589}$$

where A is now the area of the *horizon* of the black hole. It is given by $A = 4\pi R^2 = 4\pi(2GM/c^2)^2$. In particular, the result implies that every black hole has an entropy. Black

* The energy of the universe is constant. Its entropy tends towards a maximum.

holes are thus disordered systems described by thermostatics. *Black holes are the most disordered systems known.**

As an interesting note, the maximum entropy also gives a memory limit for memory chips. Can you find out how? Challenge 1314 n

What are the different microstates leading to this macroscopic entropy? It took many years to convince physicists that the microstates have to do with the various possible states of the horizon itself, and that they are due to the diffeomorphism invariance at this boundary. As Gerard 't Hooft explains, the entropy expression implies that the number of degrees of freedom of a black hole is about (but not exactly) one per Planck area of the horizon. Ref. 846 Challenge 1315 n

If black holes have entropy, they must have a temperature. What does this temperature mean? In fact, nobody believed this conclusion until two unrelated developments confirmed it within a few months. All these results were waiting to be discovered since the 1930s, even though, incredibly, nobody had thought about them for over 40 years.

Measuring acceleration with a thermometer: Fulling–Davies–Unruh radiation

Independently, Stephen Fulling in 1973, Paul Davies in 1975 and William Unruh in 1976 made the same theoretical discovery while studying quantum theory: if an inertial observer observes that he is surrounded by vacuum, a second observer *accelerated* with respect to the first does not: he observes black body radiation. The radiation has a spectrum corresponding to the temperature Ref. 847

$$T = a\frac{\hbar}{2\pi kc} \; . \qquad (590)$$

The result means that there is no vacuum on Earth, because any observer on its surface can maintain that he is accelerated with $9.8\,\mathrm{m/s^2}$, thus leading to $T = 40\,\mathrm{zK}$! We can thus measure gravity, at least in principle, using a thermometer. However, even for the largest practical accelerations the temperature values are so small that it is questionable whether the effect will ever be confirmed experimentally. But if it will, it will be a great experiment. Ref. 848

When this effect was predicted, people studied the argument from all sides. For example, it was then found that the acceleration of a *mirror* leads to radiation emission! Mirrors are thus harder to accelerate than other bodies of the same mass.

When the acceleration is high enough, also *matter* particles can be detected. If a particle counter is accelerated sufficiently strongly across the vacuum, it will start counting particles! We see that the difference between vacuum and matter becomes fuzzy at large energies.

* The precise discussion that black holes are the most disordered systems in nature is quite subtle. It is summarized by Bousso. Bousso claims that the area appearing in the maximum entropy formula cannot be taken naively as the area at a given time, and gives four arguments why this should be not allowed. However, all four arguments are wrong in some way, in particular because they assume that lengths smaller than the Planck length or larger than the universe's size can be measured. Ironically, he brushes aside some of the arguments himself later in the paper, and then deduces an improved formula, which is exactly the same as the one he criticizes first, just with a different interpretation of the area A. In short, the expression of black hole entropy is the maximum entropy for a physical system with surface A. Ref. 845

For completeness, we mention that also an observer in rotational motion detects radiation following expression (590).

Black holes aren't black

In 1974, the English physicist Stephen Hawking, famous for the courage with which he fights a disease which forces him into the wheelchair, surprised the world of general relativity with a fundamental theoretical discovery. He found that if a virtual particle–antiparticle pair appeared in the vacuum near the horizon, there is a finite chance that one particle escapes as a real particle, while the virtual antiparticle is captured by the black hole. The virtual antiparticle is thus of negative energy, and reduces the mass of the black hole. The mechanism applies both to fermions and bosons. From far away this effect looks like the emission of a particle. Hawking's detailed investigation showed that the effect is most pronounced for photon emission. In particular, Hawking showed that black holes radiate as black bodies.

Black hole radiation confirms both the result on black hole entropy by Bekenstein and the effect for observers accelerated in vacuum found by Fulling, Davies and Unruh. When all this became clear, a beautiful Gedanken experiment was published by William Unruh and Robert Wald, showing that the whole result could have been deduced already 50 years earlier! Ref. 849

Shameful as this delay of the discovery is for the community of theoretical physicists, the story itself remains beautiful. It starts in the early 1970s, when Robert Geroch studied the issue shown in Figure 330. Imagine a mirror box full of heat radiation, thus full of light. The mass of the box is assumed to be negligible, such as a box made of thin aluminium paper. We lower the box, with all its contained radiation, from a space station towards a black hole. On the space station, lowering the weight of the heat radiation allows to generate energy. Obviously, when the box reaches the black hole horizon, the heat radiation is red-shifted to infinite wavelength. At that point, the full amount of energy originally contained in the heat radiation has been provided to the space station. We can now do the following: we can open the box on the horizon, let drop out whatever is still inside, and wind the empty and massless box back up again. As a result, we have completely converted heat radiation into mechanical energy. Nothing else has changed: the black hole has the same mass as beforehand.

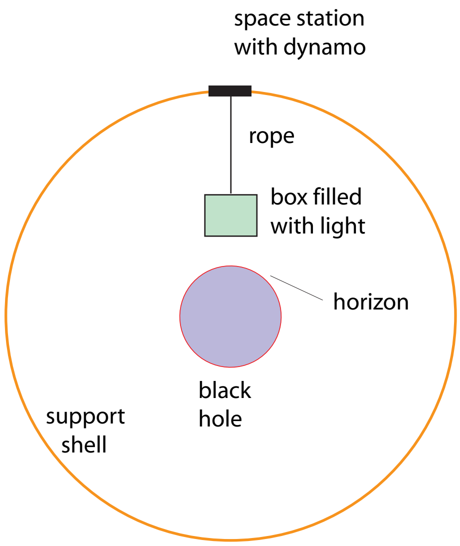


Figure 330 A *Gedanken experiment* allowing to deduce the existence of black hole radiation

But this result contradicts the second principle of thermodynamics! Geroch concluded that something must be wrong. We must have forgotten an effect which makes this process impossible.

In the 1980s, Unruh and Wald showed that black hole radiation is precisely the forgotten effect that puts everything right. Because of black hole radiation, the box feels buoyancy, so that it cannot be lowered down to the horizon. It floats somewhat above it, so that the heat radiation inside the box has not yet *zero* energy when it falls out of the opened box. As a result, the black hole does increase in mass and thus in entropy. In summary, when the empty box is pulled up again, the final situation is thus the following: only part of the energy of the heat radiation has been converted into mechanical energy, part of the energy went into the increase of mass and thus of entropy of the black hole. The second principle of thermodynamics is saved.

Well, it is only saved if the heat radiation has precisely the right energy density at the horizon and above. Let us have a look. The centre of the box can only be lowered up to a hovering distance d above the horizon, where the acceleration due to gravity is $g = c^2/4GM$. The energy E gained by lowering the box is

$$E = mc^2 - mg\frac{d}{2} = mc^2(1 - \frac{dc^2}{8GM}) \qquad (591)$$

The efficiency of the process is $\eta = E/mc^2$. To be consistent with the second law of thermodynamics, this efficiency must obey

$$\eta = \frac{E}{mc^2} = 1 - \frac{T_{BH}}{T}e \qquad (592)$$

We thus find a black hole temperature T_{BH} given by the hovering distance d. That hovering distance d is roughly given by the size of the box. The box size in turn must be at least the wavelength of the thermal radiation; in first approximation, Wien's relation gives $d \approx \hbar c/kT$. A precise calculation, first performed by Hawking, gives the result

$$T_{BH} = \frac{\hbar c^3}{8\pi kGM} = \frac{\hbar c}{4\pi k}\frac{1}{R} = \frac{\hbar}{2\pi kc}\, g_{surf} \quad \text{with} \quad g_{surf} = \frac{c^4}{4GM} \qquad (593)$$

where R and M are the radius and the mass of the black hole. It is either called the black-hole temperature or Bekenstein-Hawking temperature. As an example, a black hole with the mass of the Sun would have the rather small temperature of 62 nK, whereas a smaller black hole with the mass of a mountain, say 10^{12} kg, would have a temperature of 123 GK. That would make quite a good oven. All known black hole candidates have masses in the range from a few to a few million solar masses. The radiation is thus extremely weak, the reason being that the emitted wavelength is of the order of the black hole radius, as you might want to check. The radiation emitted by black holes is often also called *Bekenstein-Hawking radiation*.

Challenge 1316 ny

Ref. 850

Black hole radiation is thus so weak that we can speak of an academic effect. It leads to a luminosity that increases with decreasing mass or size as

Challenge 1317 ny

$$L \sim \frac{1}{M^2} \sim \frac{1}{R^2} \quad \text{or} \quad L = nA\sigma T^4 = n\frac{c^6\hbar}{G^2M^2}\frac{\pi^2}{15\cdot 2^7} \qquad (594)$$

Page 565 where σ is the Stefan–Boltzmann or *black body radiation constant*, n is the number of particle degrees of freedom that can be radiated; if only photons are radiated, we have $n = 2$. (For example, if neutrinos were massless, they would be emitted more frequently than photons.) Ref. 851

Black holes thus shine, and the more the smaller they are. This is a genuine *quantum* effect, since classically, black holes, as the name says, cannot emit any light. Even though the effect is academically weak, it will be of importance later on. In actual systems, many other effect around black holes increase the luminosity far above this value; indeed, black holes are usually brighter than normal stars, due to the radiation emitted by the matter falling into them. But that is another story. Here we are only treating isolated black holes, surrounded only by vacuum.

Due to the emitted radiation, black holes gradually lose mass. Therefore their theoretical lifetime is *finite*. A calculation shows that it is given by Challenge 1318 ny

$$t = M^3 \frac{20\,480\,\pi\, G^2}{\hbar c^4} \approx M^3\ 3.4 \cdot 10^{-16}\ \mathrm{s/kg^3} \qquad (595)$$

as function of their initial mass M. For example, a black hole with mass of 1 g would have a lifetime of $3.4 \cdot 10^{-25}$ s, whereas a black hole of the mass of the Sun, $2.0 \cdot 10^{30}$ kg, would have a lifetime of about 10^{68} years. Obviously, these numbers are purely academic. In any case, *black holes evaporate*. However, this extremely slow process for usual black holes determines their lifetime only if no other, faster process comes into play. We will present a few such processes shortly. Hawking radiation is the weakest of all known effects. It is not masked by stronger effects only if the black hole is non-rotating, electrically neutral and with no matter falling into it from the surroundings.

So far, none of these quantum gravity effects has been confirmed experimentally, as the values are much too small to be detected. However, the deduction of a Hawking temperature has been beautifully confirmed by a theoretical discovery of Unruh, who found that there are configurations of fluids in which sound waves cannot escape, so-called '*silent holes*'. Consequently, these silent holes radiate sound waves with a temperature satisfying the same formula as real black holes. A second type of analogue system, namely *optical black holes*, are also being investigated. Ref. 852 Ref. 854

Gamma ray bursts

In 1975, a much more dramatic radiation effect than black hole radiation was predicted for *charged* black holes by Damour and Ruffini. Charged black holes have a much shorter lifetime than just presented, because during their formation a second process takes place. In a region surrounding them the electric field is larger than the so-called vacuum polarization value, so that large numbers of electron-positron pairs are produced, which then almost all annihilate. This process effectively reduces the charge of the black hole to a value for which the field is below critical everywhere, while emitting large amounts of high energy light. It turns out that the mass is reduced by up to 30 % in a time of the order of seconds. That is quite shorter than 10^{68} years. This process thus produces an extremely intense *gamma ray burst*. Ref. 855

Such gamma ray bursts had been discovered in the late 1960s by military satellites which were trying to spot nuclear explosions around the world through their gamma ray emission. The satellites found about two such bursts per day, coming from all over the sky. Another satellite, the Compton satellite, confirmed that they were extragalactic in origin, and that their duration varied between a sixtieth of a second and about a thousand seconds. In 1996, the Italian-Dutch BeppoSAX satellite started mapping and measuring gamma ray bursts systematically. It discovered that they were followed by an *afterglow* in the X-ray domain of many hours, sometimes of days. In 1997 afterglow was discovered also in the optical domain. The satellite also allowed to find the corresponding X-ray, optical and radio sources for each burst. These measurements in turn allowed to determine the distance of the burst sources; red-shifts between 0.0085 and 4.5 were measured. In 1999 it also became possible to detect *optical* bursts corresponding to the gamma ray ones.*

Ref. 856
Ref. 857

All this data together show that the gamma ray bursts have energies ranging from 10^{40} W to 10^{45} W. The larger value is about one hundredth of the brightness all stars of the whole visible universe taken together! Put differently, it is the same amount of energy that is released when converting several solar masses into radiation within a few seconds. In fact, the measured luminosity is near the theoretical maximum luminosity a body can have. This limit is given by

Challenge 1319 n

$$L < L_{\text{Pl}} = \frac{c^5}{4G} = 0.9 \cdot 10^{52}\ \text{W}\ , \tag{596}$$

as you might want to check yourself. In short, the sources of gamma ray bursts are the biggest bombs found in the universe. In fact, more detailed investigations of experimental data confirm that gamma ray bursts are ‘primal screams’ of black holes in formation.

Challenge 1320 e
Ref. 857

With all this new data, Ruffini took up his 1975 model again in 1997 and with his collaborators showed that the gamma ray bursts generated by the annihilation of electron-positrons pairs created by vacuum polarization, in the region they called the *dyadosphere*, have a luminosity and a duration exactly as measured, if a black hole of about a few up to 30 solar masses is assumed. Charged black holes therefore reduce their charge and mass through the vacuum polarization and electron positron pair creation process. (The process reduces the mass because it is one of the few processes which is *reversible*; in contrast, most other attempts to reduce charge on a black hole, e.g. by throwing in a particle with the opposite charge, increase the mass of the black hole and are thus irreversible.) The left over remnant then can lose energy in various ways and also turns out to be responsible for the afterglow discovered by the BeppoSAX satellite. Among others, Ruffini’s team speculates that the remnants are the sources for the high energy cosmic rays, whose origin had not been localized so far. All these exciting studies are still ongoing.

Ref. 858

Recent studies distinguish two classes of gamma ray bursts. Short gamma ray bursts, with a duration between a millisecond and two seconds, differ significantly in energy and spectrum from long gamma ray bursts, with an average length of ten seconds, a higher energy content and a softer energy spectrum. It is often speculated that short bursts are due to merging neutron stars or merging black holes, whereas long bursts are emitted, as just explained, when a black hole is formed in a supernova or hypernova explosion. It

Ref. 860

* For more about this fascinating topic, see the http://www.aip.de/~jcg/grb.html website by Jochen Greiner.

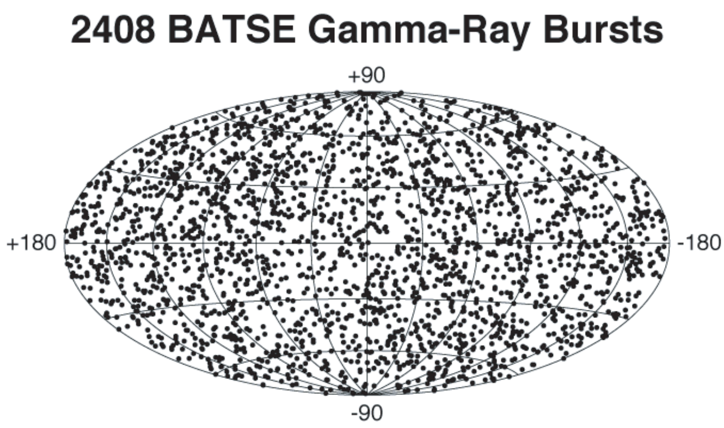


Figure 331 A selection of gamma ray bursters observed in the sky

also seems that gamma ray bursts are not of spherical symmetry, but that the emission takes place in a collimated beam. This puts the energy estimates given above somewhat into question. The details of the formation process are still subject to intense exploration.

Other processes leading to emission of radiation from black holes are also possible. Examples are matter falling into the black hole and heating up, matter being ejected from Page 448 rotating black holes through the Penrose process, or charged particles falling into a black hole. These mechanisms are at the origin of *quasars*, the extremely bright quasi-stellar sources found all over the sky. They are assumed to be black holes surrounded by matter, in the development stage following gamma ray bursters. The details of what happens in quasars, the enormous voltages (up to 10^{20} V) and magnetic fields generated, as well as their effects on the surrounding matter are still object of intense research in astrophysics.

Material properties of black holes

Once the concept of entropy of a black hole was established, people started to think about black holes like about any other material object. For example, black holes have a matter density, which can be defined by relating their mass to a fictitious volume defined by $4\pi R^3/3$. This density is given by

$$\rho = \frac{1}{M^2} \frac{3c^6}{32\pi G^3} \tag{597}$$

and can be quite low for large black holes. For the highest black holes known, with 1000 million solar masses or more, the density is of the order of the density of air. By the way, the gravitational acceleration at the horizon is still appreciable, as it is given by

$$g_{\text{surf}} = \frac{1}{M} \frac{c^4}{4G} = \frac{c^2}{2R} \tag{598}$$

Challenge 1321 ny which is still $15\,\text{km/s}^2$ for an air density black hole.

Obviously, the black hole temperature is related to the entropy S by its usual definition

$$\frac{1}{T} = \left.\frac{\partial S}{\partial E}\right|_{\rho} = \left.\frac{\partial S}{\partial (Mc^2)}\right|_{\rho} \tag{599}$$

All other thermal properties can be deduced by the standard relations from thermostatics.

In particular, it looks as if black holes are the matter states with the largest possible entropy. Can you confirm this statement?

Challenge 1322 ny

It also turns out that black holes have a *negative* heat capacity: when heat is added, they cool down. In other words, black holes cannot achieve equilibrium with a bath. This is not a real surprise, since *any* gravitationally bound material system has negative specific heat. Indeed, it takes only a bit of thinking to see that any gas or matter system collapsing under gravity follows $dE/dR > 0$ and $dS/dR > 0$. That means that while collapsing, the energy and the entropy of the system shrink. (Can you find out where they go?) Since temperature is defined as $1/T = dS/dE$, temperature is always positive; from the temperature increase $dT/dR < 0$ during collapse one deduces that the specific heat dE/dT is negative.

Challenge 1323 ny
Challenge 1324 n
Ref. 861

Black holes, like any object, oscillate when slightly perturbed. These vibrations have also been studied; their frequency is proportional to the mass of the black hole.

Ref. 862

Nonrotating black holes have no magnetic field, as was established already in the 1960s by Russian physicists. On the other hand, black holes have something akin to a finite electrical conductivity and a finite viscosity. Some of these properties can be understood if the horizon is described as a membrane, even though this model is not always applicable. In any case, one can study and describe macroscopic black holes like any other macroscopic material body. The topic is not closed.

Ref. 851
Ref. 859

How do black holes evaporate?

When a nonrotating and uncharged black hole loses mass by radiating Hawking radiation, eventually its mass reaches values approaching the Planck mass, namely a few micrograms. Expression (595) for the lifetime, applied to a black hole of Planck mass, yields a value of over sixty thousand Planck times. A surprising large value. What happens in those last instants of evaporation?

A black hole approaching the Planck mass at some time will get smaller than its own Compton wavelength; that means that it behaves like an elementary particle, and in particular, that quantum effects have to be taken into account. It is still unknown how these final evaporation steps take place, whether the mass continues to diminish smoothly or in steps (e.g. with mass values decreasing as $\sqrt{n}$ when n approaches zero), how its internal structure changes, whether a stationary black hole starts to rotate (as the author predicts), how the emitted radiation deviates from black body radiation. There is still enough to study. However, one important issue *has* been settled.

The information paradox of black holes

When the thermal radiation of black holes was discovered, one question was hotly debated for many years. The matter forming a black hole can contain lots of information;

e.g., imagine the black hole formed by a large number of books collapsing onto each other. On the other hand, a black hole radiates thermally until it evaporates. Since thermal radiation carries no information, it seems that information somehow disappears, or equivalently, that entropy increases.

An incredible number of papers have been written about this problem, some even claiming that this example shows that physics as we know it is incorrect and needs to be changed. As usual, to settle the issue, we need to look at it with precision, laying all prejudice aside. Three intermediate questions can help us finding the answer.

- What happens when a book is thrown into the Sun? When and how is the information radiated away?
- How precise is the sentence that black hole radiate *thermal* radiation? Could there be a slight deviation?
- Could the deviation be measured? In what way would black holes radiate information?

Challenge 1325 e You might want to make up your own mind before reading on.

Let us walk through a short summary. When a book or any other highly complex – or low entropy – object is thrown into the Sun, the information contained is radiated away. The information is contained in some slight deviations from black hole radiation, namely in slight correlations between the emitted radiation emitted over the burning time of the Sun. A short calculation, comparing the entropy of a room temperature book and the information contained in it, shows that these effects are extremely small and difficult to measure.

Ref. 863 A clear exposition of the topic was given by Don Page. He calculated what information would be measured in the radiation if the system of black hole and radiation *together* would be in a *pure* state, i.e. a state containing specific information. The result is simple. Even if a system is large – consisting of many degrees of freedom – and in pure state, any smaller *sub*system nevertheless looks almost perfectly thermal. More specifically, if a total system has a Hilbert space dimension $N = nm$, where n and $m \leqslant n$ are the dimensions of two subsystems, and if the total system is in a pure state, the subsystem m would have an entropy S_m given by (Challenge 1326 ny)

$$S_m = \frac{1-m}{2n} + \sum_{k=n+1}^{mn} \frac{1}{k} \tag{600}$$

which is approximately given by

$$S_m = \ln m - \frac{m}{2n} \quad \text{for} \quad m \gg 1 . \tag{601}$$

To discuss the result, let us think of n and m as counting degrees of freedom, instead of Hilbert space dimensions. The first term in equation (601) is the usual entropy of a mixed state. The second term is a small deviation and describes the amount of specific information contained in the original pure state; inserting numbers, one finds that it is extremely small compared to the first. In other words, the subsystem m is almost indistinguishable from a mixed state; it looks like a thermal system even though it is not.

A calculation shows that the second, small term on the right of equation (601) is indeed sufficient to radiate away, during the lifetime of the black hole, any information contained in it. Page then goes on to show that the second term is so small that not only it is lost in measurements; it is also lost in the usual, perturbative calculations for physical systems.

The question whether any radiated information could be measured can now be answered directly. As Don Page showed, even measuring half of the system only gives Ref. 864 about 1/2 bit of that information. It is necessary to measure the complete system to measure all the contained information. In summary, at a given instant, the amount of information radiated by a black hole is negligible when compared with the total black hole radiation, and is practically impossible to detect by measurements or even by usual calculations.

More paradoxes

A black hole is a macroscopic object, similar to a star. Like all objects, it can interact with its environment. It has the special property to swallow everything that falls into them. This immediately leads us to ask if we can use this property to cheat around the usual everyday 'laws' of nature. Some attempts have been studied in the section on general Page 449 relativity and above; here we explore a few additional ones.

▪ Apart from the questions of entropy, we can look for methods to cheat around con- Challenge 1327 ny servation of energy, angular momentum, or charge. Every Gedanken experiment comes to the same conclusions. No cheats are possible; in addition, the maximum number of degrees of freedom in a region is proportional to the surface area of the region, and *not* to its volume. This intriguing result will keep us busy for quite some time.

▪ A black hole transforms matter into antimatter with a certain efficiency. Thus one might look for departures from particle number conservation. Are you able to find an Challenge 1328 ny example?

▪ Black holes deflect light. Is the effect polarization dependent? Gravity itself makes no difference of polarization; however, if virtual particle effects of QED are included, the story Ref. 868 might change. First calculations seem to show that such a effect exists, so that gravitation might produce rainbows. Stay tuned.

▪ If lightweight boxes made of mirrors can float in radiation, one gets a strange consequence: such a box might self-accelerate in free space. In a sense, an accelerated box could float on the Fulling–Davies–Unruh radiation it creates by its own acceleration.

Are you able to show the following: one reason why this is impossible is a small but Challenge 1329 ny difference between gravity and acceleration, namely the absence of tidal effects. (Other reasons, such as the lack of perfect mirrors, also make the effect impossible.)

▪ In 2003, Michael Kuchiev has made the spectacular prediction that matter and radiation with a wavelength *larger* than the diameter of a black hole is partly reflected when Ref. 869 it hits a black hole. The longer the wavelength, the more efficient the reflection would be. For stellar or even bigger black holes, only photons or gravitons are predicted to be reflected. Black holes are thus not complete trash cans. Is the effect real? The discussion is still ongoing.

Quantum mechanics of gravitation

Let us take a conceptual step at this stage. So far, we looked at quantum theory *with* gravitation; now we have a glimpse at quantum theory *of* gravitation.

If we focus on the similarity between the electromagnetic field and the gravitational 'field,' we should try to find the quantum description of the latter. Despite attempts by many brilliant minds for almost a century, this approach was not successful.* Let us see why.

The gravitational Bohr atom

Challenge 1330 ny

A short calculation shows that an electron circling a proton due to gravity alone, without electrostatic attraction, would do so at a gravitational Bohr radius of

$$r_{\mathrm{gr.B.}} = \frac{\hbar^2}{G\, m_\mathrm{e}^2\, m_\mathrm{p}} = 1.1 \cdot 10^{29}\ \mathrm{m} \tag{602}$$

Challenge 1331 ny

which is about a thousand times the distance to the cosmic horizon. In fact, even in the normal hydrogen atom there is not a *single way* to measure gravitational effects. (Are you able to confirm this?) But why is gravity so weak? Or equivalently, why are the universe and normal atoms so much smaller than a gravitational Bohr atom? At the present point of our quest these questions cannot be answered. Worse, the weakness of gravity even means that with high probability, future experiments will provide little additional data helping to decide among competing answers. The only help is careful thought.

Decoherence of space-time

If the gravitational field evolves like a quantum system, we encounter all issues found in other quantum systems. General relativity taught us that the gravitational field and space-time are the same. As a result, we may ask why no superpositions of different macroscopic space-times are observed.

Ref. 870
Page 742
Challenge 1332 ny

The discussion is simplified for the simplest case of all, namely the superposition, in a vacuum region of size l, of a homogeneous gravitational field with value g and one with value g'. As in the case of a superposition of macroscopic distinct wavefunctions, such a superposition *decays*. In particular, it decays when particles cross the volume. A short calculation yields a decay time given by

$$t_\mathrm{d} = \left(\frac{2kT}{\pi m}\right)^{3/2} \frac{n l^4}{(g - g')^2}\ , \tag{603}$$

Challenge 1333 ny

where n is the particle number density, kT their kinetic energy and m their mass. Inserting typical numbers, we find that the variations in gravitational field strength are *extremely* small. In fact, the numbers are so small that we can deduce that the gravitational field is the *first* variable which behaves classically in the history of the universe. Quantum gravity effects for space-time will thus be extremely hard to detect.

In short, matter not only tells space-time how to curve, it also tells it to behave with class. This result calls for the following question.

* Modern approaches take another direction, as explained in the third part of the mountain ascent.

Do gravitons exist?

Quantum theory says that everything that moves is made of particles. What kind of particles are gravitational waves made of? If the gravitational field is to be treated quantum mechanically like the electromagnetic field, its waves should be quantized. Most properties of these quanta can be derived in a straightforward way.

The $1/r^2$ dependence of universal gravity, like that of electricity, implies that the particles have *vanishing* mass and move at light speed. The independence of gravity from electromagnetic effects implies a *vanishing* electric charge.

The observation that gravity is always attractive, never repulsive, means that the field quanta have *integer* and *even* spin. Vanishing spin is ruled out, since it implies no coupling to energy. To comply with the property that 'all energy has gravity', $S = 2$ is needed. In fact, it can be shown that *only* the exchange of a massless spin 2 particle leads, in the classical limit, to general relativity. Ref. 871

The coupling strength of gravity, corresponding to the fine structure constant of electromagnetism, is given either by

$$\alpha_{G1} = \frac{G}{\hbar c} = 2.2 \cdot 10^{-15}\,\mathrm{kg}^{-2} \quad \text{or by} \quad \alpha_{G2} = \frac{Gmm}{\hbar c} = \left(\frac{m}{m_{Pl}}\right)^2 = \left(\frac{E}{E_{Pl}}\right)^2 \qquad (604)$$

However, the first expression is not a pure number; the second expression is, but depends on the mass one inserts. These difficulties reflect the fact that gravity is not properly speaking an interaction, as became clear in the section on general relativity. It is often argued that m should be taken as the value corresponding to the energy of the system in question. For everyday life, typical energies are 1 eV, leading to a value $\alpha_{G2} \approx 1/10^{56}$. Gravity is indeed weak compared to electromagnetism, for which $\alpha_{em} = 1/137.04$.

If all this is correct, *virtual* field quanta would also have to exist, to explain static gravitational fields.

However, up to this day, the so-called *graviton* has not yet been detected, and there is in fact little hope that it ever will. On the experimental side, nobody knows yet how to build a graviton detector. Just try! On the theoretical side, the problems with the coupling constant probably make it impossible to construct a *renormalisable* theory of gravity; the lack of renormalization means the impossibility to define a perturbation expansion, and thus to define particles, including the graviton. It might thus be that relations such as $E = \hbar\omega$ or $p = \hbar/2\pi\lambda$ are not applicable to gravitational waves. In short, it may be that the particle concept has to be changed before applying quantum theory to gravity. The issue is still open at this point. Challenge 1334 n

Space-time foam

The indeterminacy relation for momentum and position also applies to the gravitational field. As a result, it leads to an expression for the indeterminacy of the metric tensor g in a region of size L, which is given by

$$\Delta g \approx 2\frac{l_{Pl}{}^2}{L^2}\,, \qquad (605)$$

Challenge 1335 ny where $l_{Pl} = \sqrt{\hbar G/c^3}$ is the Planck length. Can you deduce the result? Quantum theory thus shows that like the momentum or the position of a particle, also the metric tensor g is a *fuzzy* observable.

But that is not all. Quantum theory is based on the principle that actions below $\hbar/2$ cannot be observed. This implies that the observable values for the metric g in a region of size L are bound by

$$g \geqslant \frac{2\hbar G}{c^3}\frac{1}{L^2} \ . \qquad (606)$$

Can you confirm this? The result has far-reaching consequences. A minimum value for the metric depending inversely on the region size implies that it is impossible to say what happens to the shape of space-time at extremely small dimensions. In other words, at extremely high energies, the concept of space-time itself becomes fuzzy. John Wheeler introduced the term *space-time foam* to describe this situation. The term makes clear that space-time is not continuous nor a manifold in those domains. But this was the basis on which we built our description of nature so far! We are forced to deduce that our description of nature is built on sand. This issue will form the start of the third part of Page 920 our mountain ascent.

No particles

Gravity has another important consequence for quantum theory. To count and define Page 721 particles, quantum theory needs a defined vacuum state. However, the vacuum state cannot be defined when the curvature radius of space-time, instead of being larger than the Compton wavelength, becomes comparable to it. In such highly curved space-times, particles cannot be defined. The reason is the impossibility to distinguish the environment from the particle in these situations: in the presence of strong curvatures, the vacuum is full of spontaneously generated matter, as black holes show. Now we just saw that at small dimensions, space-time fluctuates wildly; in other words, space-time is highly curved at small dimensions or high energies. In other words, strictly speaking particles cannot be defined; the particle concept is only a low energy approximation! We will explore this strange conclusion in more detail in the third part of our mountain ascent.

No science fiction

The end of the twentieth century has brought several unexpected but strong results in the semiclassical quantum gravity.

Ref. 872 In 1995 Ford and Roman found that worm holes, which are imaginable in general relativity, cannot exist if quantum effects are taken into account. They showed that macroscopic worm holes require unrealistically large negative energies. (For microscopic worm holes the issue is still unclear.)

Ref. 874 In 1996 it was found by Kay, Radzikowski and Wald that closed time-like curves do not exist in semiclassical gravity; there are thus no time machines in nature.

Ref. 873 In 1997 Pfenning and Ford showed that warp drive situations, which are also imaginable in general relativity, cannot exist if quantum effects are taken into account. They also require unrealistically large negative energies.

Not cheating any longer

This short excursion into the theory of quantum gravity showed that a lot of trouble is waiting. The reason is that up to now, we deluded ourselves. In fact, it was more than that: we cheated. We carefully hid a simple fact: quantum theory and general relativity *contradict* each other. That was the real reason that we stepped back to special relativity before we started exploring quantum theory. In this way we avoided all problems, as quantum theory does not contradict *special* relativity. However, it does contradict *general* relativity. The issues are so dramatic, changing everything from the basis of classical physics to the results of quantum theory, that we devote the beginning of the third part only to the exploration of the contradictions. There will be surprising consequences on the nature of space-time, particles and motion. But before we study these issues, we complete the theme of the present, second part of the mountain ascent, namely the essence of matter and interactions.

Bibliography

728 Gell-Mann wrote this for the 1976 Nobel Conference. M. Gell-Mann, What are the Building Blocks of Matter?, in D. Huff & O. Prewitt, editors, *The Nature of the Physical Universe*, New York, Wiley, 1979, p. 29. Cited on page 739.

729 See e.g. the reprints of his papers in the standard collection by John A. Wheeler & Wojciech H. Zurek, *Quantum Theory and Measurement*, Princeton University Press, 1983. Cited on page 740.

730 H.D. Zeh, On the interpretation of measurement in quantum theory, *Foundations of Physics* **1**, pp. 69–76, 1970. Cited on page 740.

731 Linda Reichl, *A Modern Course in Statistical Physics*, Wiley, 2nd edition, 1998. An excellent introduction into thermodynamics. Cited on page 741.

732 E. Joos & H.D. Zeh, The emergence of classical properties through interactions with the environment, *Zeitschrift für Physik B* **59**, pp. 223–2443, 1985. Cited on page 742.

733 M. Tegmark, Apparent wave function collapse, Foundation of Physics Letters, **6**, pp. 571–590, 1993. Cited on page 742.

734 The decoherence time is bound from above by the relaxation time. See A.O. Caldeira & A.J. Leggett, Influence of damping on quantum interference: an exactly soluble model, *Physical Review A* **31**, 1985, pp. 1059–1066. This is the main reference about effects of decoherence for a harmonic oscillator. The general approach to relate decoherence to the influence of the environment is due to Niels Bohr, and has been pursued in detail by Hans Dieter Zeh. Cited on page 744.

735 G. Lindblad, On the generators of quantum dynamical subgroups, *Communications in Mathematical Physics* **48**, pp. 119–130, 1976. Cited on page 744.

736 Wojciech H. Zurek, Decoherence and the transition from quantum to classical, *Physics Today* pp. 36–44, October 1991. An easy but somewhat confusing article. His reply to the numerous letters of response in *Physics Today, April* 1993, pp. 13–15, and pp. 81–90, exposes his ideas in a clearer way and gives a taste of the heated discussions on this topic. Cited on pages 745 and 749.

737 John Bardeen, his talk during his trip to Japan. Cited on page 745.

738 The first decoherence measurement was performed by M. Brune & al., Observing progressive decoherence of the "meter" in a quantum measurement, *Physical Review Letters* 77, pp. 4887–4890, 9 December 1996. Cited on page 745.

739 Other experiments confirming the numerical predictions from decoherence were published by C. Monroe & al., *Science* **272**, p. 1131, 1996, W.P. Schleich & al., *Nature* **403**, p. 256, 2000, C.J. Myatt & al., *Nature* **403**, p. 269, 2000. See also the summary by W.T. Strunz, G. Alber & F. Haake, Dekohärenz in offenen Quantensystemen, *Physik Journal* **1**, pp. 47–52, November 2002. Cited on page 745.

740 K. Baumann, Quantenmechanik und Objektivierbarkeit, *Zeitschrift für Naturforschung* **25a**, pp. 1954–1956, 1970. Cited on page 746.

741 See for example D. Styer, *Physics Today* p. 11, September 2000. Cited on page 747.

742 David Bohm, *Quantum Theory*, Prentice-Hall, 1951, pp. 614–622. Cited on page 747.

743 Albert Einstein, Boris Podolsky & Nathan Rosen, Can quantum-mechanical description of reality be considered complete?, *Physical Review* **48**, pp. 696–702, 1935. Cited on page 747.

744 Alain Aspect, J. Dalibard & G. Roger, Experimental tests of Bell's inequalities using time-varying analysers, *Physical Review Letters* **49**, pp. 1804–1807, 1982, Cited on page 748.

745 John Maddox, ... Cited on page 748.

746 David Z. Albert, *Quantum Mechanics and Experience*, Harvard University Press, 1994. No citations.

747 On the superposition of magnetization in up and down directions there are numerous papers. Recent experiments on the subject of quantum tunnelling in magnetic systems are described in D.D. Awschalom, J.F. Smith, G. Grinstein, D.P. DiVicenzo & D. Loss, Macroscopic quantum tunnelling in magnetic proteins, *Physical Review Letters* **88**, pp. 3092–3095, 1992, and in C. Paulsen & al., Macroscopic quantum tunnelling effects of Bloch walls in small ferromagnetic particles, *Europhysics Letters* **19**, pp. 643–648, 1992. Cited on page 749.

748 The prediction that quantum tunnelling could be observable when the dissipative interaction with the rest of the world is small enough was made by Leggett; the topic is reviewed in A.J. Leggett, S. Chahravarty, A.T. Dorsey, M.P.A. Fisher, A. Garg & W. Zwerger, Dynamics of dissipative 2-state systems, *Review of Modern Physics* **59**, pp. 1–85, 1987. Cited on page 752.

749 For example, superpositions were observed in Josephson junctions by R.F. Voss & R.A. Webb, Macroscopic quantum tunnelling in 1 mm Nb Josephson junctions, *Physical Review Letters* **47**, pp. 265–268, 1981, Cited on page 749.

750 S. Haroche, Entanglement, decoherence and the quantum-classical transition, *Physics Today* **51**, pp. 36–42, July 1998. An experiment putting atom at two places at once, distant about 80 nm, was published by C. Monroe & al., *Science* **272**, p. 1131, 1996. Cited on page 749.

751 M.R. Andres & al., Observations of interference between two Bose condensates, *Science* **275**, pp. 637–641, 31 January 1997. See also the www.aip.org/physnews/special.htm website. Cited on page 749.

752 The most famous reference on the wave function collapse is chapter IV of the book by Kurt Gottfried, *Quantum Mechanics*, Benjamin, New York, 1966, It is the favorite reference by Victor Weisskopf, and cited by him on every occasion he talks about the topic. Cited on page 750.

753 The first experimental measurement of superpositions of left and right flowing currents is ... In the year 2000, superpositions of 1 μA clockwise and anticlockwise have been detected; for more details, see J.R. Friedman & al., Quantum superposition of distinct macroscopic states, *Nature* **406**, p. 43, 2000. Cited on page 749.

754 Collapse times have been measured ... Cited on page 754.

755 A clear discussion can be found in S. Haroche & J.-M. Raimond, Quantum computing: dream or nightmare?, *Physics Today* **49**, pp. 51–52, 1996, as well as the comments in *Physics Today* **49**, pp. 107–108, 1996. Cited on page 749.

756 S. Kochen & E.P. Specker, The problem of hidden variables in quantum mechanics, **17**, pp. 59–87, 1967. Cited on page 754.

757 J.F. Clauser, M.A. Horne, A. Shimony & R.A. Holt, *Physical Review Letters* **23**, p. 880, 1969. The more general and original result is found in J.S. Bell, *Physics* **1**, p. 195, 1964, Cited on page 755.

758 Bryce de Witt & al., *The Many–Worlds Interpretation of Quantum Mechanics*, Princeton University Press, 1973. This interpretation talks about entities which cannot be observed, namely the many worlds, and often assumes that the wavefunction of the universe exists. Both habits are beliefs and in contrast with facts. Cited on page 757.

759 'On the other had I think I can safely say that nobody understands quantum mechanics." From Richard P. Feynman, *The Character of Physical Law*, MIT Press, Cambridge, 1965, p. 129. He repeatedly made this statement, e.g. in the introduction of his otherwise excellent *QED: the Strange Theory of Light and Matter*, Penguin Books, 1990.. Cited on page 758.

760 M. Tegmark, The importance of quantum decoherence in brain processes, *Physical Review D* **61**, pp. 4194–4206, 2000, or also http://www.arxiv.org/abs/quant-ph/9907009. Cited on page 758.

761 Connections between quantum theory and information theory can be found in ... Cited on page 759.

762 J.A. Wheeler, pp. 242–307, in *Batelle Recontres: 1967 Lectures in Mathematics and Physics*, C. DeWitt & J.A. Wheeler, editors, W.A. Benjamin, 1968. For a pedagogical explanation, see John W. Norbury, From Newton's laws to the Wheeler-DeWitt equation, http://www.arxiv.org/abs/physics/980604 or *European Journal of Physics* **19**, pp. 143–150, 1998. Cited on page 760.

763 E. Fermi, Quantum theory of radiation, *Review of Modern Physics* **4**, p. 87, 1932. No citations.

764 G. Hegerfeldt, Causality problems for Fermi's two-atom system, *Physical Review Letters* 72, pp. 596–599, 1994, No citations.

765 The beautiful analogy between optics and two-dimensional electron systems, in GaAs at helium temperature, has been studied in detail in C.W.J. Beenakker & H. van Houten, Solid State Physics, **44**, pp. 1–228, 1991, No citations.

766 The use of radioactivity for breeding is described by ... Cited on page 763.

767 S.M. Block, Real engines of creation, *Nature* **386**, pp. 217–219, 1997. Cited on page 764.

768 Early results and ideas on molecular motors are summarised by Barbara Goss Levi, Measured steps advance the understanding of molecular motors, *Physics Today* pp. 17–19,

April 1995. Newer results are described in R.D. Astumian, Making molecules into motors, *Scientific American* pp. 57–64, July 2001. Cited on page 765.

769 The motorized screw used by viruses was described by A.A. Simpson & al., Structure of the bacteriophage phi29 DNA packaging motor, *Nature* **408**, pp. 745–750, 2000. Cited on page 764.

770 R. Bartussek & P. Hänggi, Brownsche Motoren, *Physikalische Blätter* **51**, pp. 506–507, 1995. Cited on page 765.

771 The table and the evolutionary tree are taken from J.O. McInerney, M. Mullarkey, M.E. Wernecke & R. Powell, Bacteria and Archaea: Molecular techniques reveal astonishing diversity, *Biodiversity* 3, pp. 3–10, 2002. Cited on page 767.

772 This is taken from the delightful children text Hans J. Press, *Spiel das Wissen schafft*, Ravensburger Buchverlag 1964, 2004. Cited on page 769.

773 Linda Buck and Richard Axel received the 2004 Nobel prize for medicine and physiology for their unravelling of the working of the sense of smell. Cited on page 770.

774 To learn to enjoy life to the maximum there is no standard way. A good foundation can be found in those books which teach the ability to those which have lost it. Cited on page 770.

The best experts are those who help others to overcome traumas. Peter A. Levine & Ann Frederick, *Waking the Tiger – Healing Trauma – The Innate Capacity to Transform Overwhelming Experiences*, North Atlantic Books, 1997. Geoff Graham, *How to Become the Parent You Never Had - a Treatment for Extremes of Fear, Anger and Guilt*, Real Options Press, 1986.

A good complement to these texts is the systemic approach, as presented by Bert Hellinger, *Zweierlei Glück*, Carl Auer Verlag, 1997. Some of his books are also available in English. The author explains how to achieve a life of contentness, living with the highest possible responsibility of one's actions, by reducing entaglements with one's past and with other people. He presents a simple and efficient technique for realising this disentanglement.

The next step, namely full mastery in the enjoyment of life, can be found in any book written by somebody who has achieved mastery in any one topic. The topic itself is not important, only the passion is.

A. de la Garanderie, *Le dialogue pédagogique avec l'élève*, Centurion, 1984, A. de la Garanderie/ *Pour une pédagogie de l'intelligence*, Centurion, 1990, A. de la Garanderie/ *Réussir ça s'apprend*, Bayard, 1994. De la Garanderie explains how the results of teaching and learning depend in particular on the importance of evocation, imagination and motivation.

Plato, *Phaedrus*, Athens, 380 BCE.

Françoise Dolto, *La cause des enfants*, Laffont, 1985, and her other books. Dolto (1908–1988), a child psychiatrist, is one of the world experts on the growth of the child; her main theme was that growth is only possible by giving the highest possible responsibility to every child during its development.

In the domain of art, many had the passion to achieve full pleasure. A good piece of music, a beautiful painting, an expressive statue or a good movie can show it. On a smaller scale, the art to typeset beautiful books, so different from what many computer programs do by default, the best introduction is by Jan Tschichold (1902–1974), the undisputed master of the field. Among the many books he designed are the beautiful Penguin books of the late 1940s; he also was a type designer, e.g. of the *Sabon* typeface. A beautiful summary of his views is the short but condensed text Jan Tschichold, *Ausgewählte Aufsätze über Fragen der Gestalt des Buches und der Typographie*, Birkhäuser Verlag, Basel, 1993. An extensive and beautiful textbook on the topic is Hans Peter Willberg & Friedrich Forssman, Lesetypographie, Verlag Hermann Schmidt, Mainz, 1997. See also ROBERT BRINGHURST, *The Elements of Typographic Style*, Hartley & Marks, 2004.

Many scientists passionately enjoyed their occupation. Any biography of Charles Darwin will purvey his fascination for biology, of Friedrich Bessel for astronomy, of Albert Einstein for physics and of Linus Pauling for chemistry.

775 The group of John Wearden in Manchester has shown by experiments with persons that the accuracy of a few per cent is possible for any action with a duration between a tenth of a second and a minute. See J. MCCRONE, When a second lasts forever, *New Scientist* pp. 53–56, 1 November 1997. Cited on page 771.

776 E.J. ZIMMERMAN, The macroscopic nature of space-time, *American Journal of Physics* **30**, pp. 97–105, 1962. Cited on page 773.

777 The potential of single-ion clocks is explained in ... Cited on page 773.

778 P.D. PEŞ IĆ, The smallest clock, *European Journal of Physics* **14**, pp. 90–92, 1993. Cited on page 773.

779 D.J MORRÉ & al., Biochemical basis for the biological clock, *Biochemistry* **41**, pp. 11941–11945, 2002. Cited on page 774.

780 The chemical clocks in our body are described in JOHN D. PALMER, *The Living Clock*, Oxford University Press, 2002, or in A. AHLGREN & F. HALBERG, *Cycles of Nature: An Introduction to Biological Rhythms*, National Science Teachers Association, Washington, 1990. See also the http://www.msi.umn.edu/~halberg/introd/ web site. Cited on page 774.

781 An introduction to the sense of time as result of clocks in the brain is found in R.B. IVRY & R. SPENCER, The neural representation of time, *Current Opinion in Neurobiology* **14**, pp. 225–232, 2004. The interval timer is explain in simple words in K. WRIGHT, Times in our lives, *Scientific American* pp. 40–47, September 2002. The MRI research used is S.M. RAO, A.R. MAYER & D.L. HARRINGTON, The evolution of brain activation during temporal processing, *Nature Neuroscience* **4**, pp. 317–323, 2001. Cited on page 774.

782 See for example ... Cited on page 776.

783 A pretty example of a quantum mechanical system showing exponential behaviour at all times is given by H. NAKAZATO, M. NAMIKI & S. PASCAZIO, Exponential behaviour of a quantum system in a macroscopic medium, *Physical Review Letters* **73**, pp. 1063–1066, 1994. Cited on page 776.

784 See the delightful book about the topic by PAOLO FACCHI & SAVERIO PASCAZIO, *La regola d'oro di Fermi*, Bibliopolis, 1999. An experiment observing deviations at short times is S.R. WILKINSON, C.F. BHARUCHA, M.C. FISCHER, K.W. MADISON, P.R. MORROW, Q. NIU, B. SUNDARAM & M.G. RAIZEN, *Nature* **387**, p. 575, 1997. Cited on page 776.

785 See ... Cited on page 777.

786 H. KOBAYASHI & S. KOHSHIMA, Unique morphology of the human eye, *Nature* **387**, pp. 767–768, 1997. Cited on page 780.

787 JAMES W. PRESCOTT, Body pleasure and the origins of violence, *The Futurist* Bethseda, 1975, also available at http://www.violence.de/prescott/bullettin/article.html. Cited on page 780.

788 See for example P. PYYKKÖ, Relativity, gold, closed-shell interactions, and CsAu.NH3, *Angewandte Chemie, International Edition* **41**, pp. 3573–3578, 2002, or L.J. NORRBY, Why is mercury liquid? Or, why do relativistic effects not get into chemistry textbooks?, *Journal of Chemical Education* **68**, pp. 110–113, 1991. Cited on page 780.

789 The switchable mirror effect was discovered by J.N. Huiberts, R. Griessen, J.H. Rector, R.J. Wijngarden, J.P. Dekker, D.G. de Groot & N.J. Koeman, Yttrium and lanthanum hydride films with switchable optical properties, *Nature* **380**, pp. 231–234, 1996. A good introduction is R. Griessen, Schaltbare Spiegel aus Metallhydriden, *Physikalische Blätter* **53**, pp. 1207–1209, 1997. Cited on page 785.

790 The present attempts to build cheap THz wave sensors – which allow to see through clothes – is described by D. Clery, Brainstorming their way to an imaging revolution, *Science* **297**, pp. 761–763, 2002. Cited on page 786.

791 V.F. Weisskopf, Of atoms, mountains and stars: a study in qualitative physics, *Science* **187**, p. 605, 1975. Cited on page 787.

792 K. Autumn, Y. Liang, T. Hsich, W. Zwaxh, W.P. Chan, T. Kenny, R. Fearing & R.J. Full, Adhesive force of a single gecko foot-hair, *Nature* **405**, pp. 681–685, 2000. Cited on page 788.

793 A.B. Kesel, A. Martin & T. Seidl, Getting a grip on spider attachment: an AFM approach to microstructure adhesion in arthropods, *Smart Materials and Structures* **13**, pp. 512–518, 2004. Cited on page 788.

794 H. Inaba, T. Saitou, K.-I. Tozaki & H. Hayashi, Effect of the magnetic field on the melting transition of H2O and D2O measured by a high resolution and supersensitive differential scanning calorimeter, *Journal of Applied Physics* **96**, pp. 6127–6132, 2004. Cited on page 788.

795 R. Hemley & al., presented these results at the International Conference on New Diamond Science in Tsukuba (Japan) 2005. Cited on page 789.

796 On the internet, the 'spherical' periodic table is credited to Timothy Stowe; but there is no reference for that claim, except an obscure calendar from a small chemical company. The table (containing a few errors) is from the http://chemlab.pc.maricopa.edu/periodic/stowetable.html website. Cited on page 780.

797 For good figures of atomic orbitals, take any modern chemistry text. See for example the text by ... Cited on page 782.

798 For experimentally determined pictures of the orbitals of dangling bonds, see for example F. Giessibl & al., Subatomic features on the silicon (111)-(7x7) surface observed by atomic force microscopy, *Science* **289**, pp. 422–425, 2000. Cited on page 782.

799 A.P. Finne, T. Araki, R. Blaauwgeers, V.B. Eltsov, N.B. Kopnin, M. Krusius, L. Skrbek, M. Tsubota & G.E. Volovik, An intrinsic velocity-independent criterion for superfluid turbulence, *Nature* **424**, pp. 1022–1025, 2003. Cited on page 791.

800 J.L. Costa–Krämer, N. Garcia, P. Garcia–Molchales & P.A. Serena, Conductance in disordered nanowires: foward and backscattering, *Physical Review B* **53**, p. 102658, 1996. Cited on page 791.

801 The first experimental evidence is by V.J. Goldman & B. Su, Resonance tunnelling in the fractional quantum Hall regime: measurement of fractional charge, *Science* **267**, pp. 1010–1012, 1995. The first unambiguous measurements are by R. de Picciotto & al., Direct observation of a fractional charge, *Nature* **389**, pp. 162–164, 1997, and L. Saminadayar, D.C. Glattli, Y. Jin & B. Etienne, Observation of the e,3 fractionally charged Laughlin quasiparticle/ *Physical Review Letters* **79**, pp. 2526–2529, 1997. or http://www.arxiv.org/abs/cond-mat/9706307. Cited on page 791.

802 The original prediction, which earned him a Nobel prize for physics, is in R.B. Laughlin, Anomalous quantum Hall effect: an incompressible quantum fluid with fractionally charged excitations, *Physical Review Letters* **50**, pp. 1395–1398, 1983. Cited on pages 791 and 792.

803 The fractional quantum Hall effect was discovered by D.C. Tsui, H.L. Stormer & A.C. Gossard, Two-dimensional magnetotransport in the extreme quantum limit, *Physical Review Letters* **48**, pp. 1559–1562, 1982. Cited on page 792.

804 The discovery of the original quantum Hall effect, which earned von Klitzing a Nobel prize, was published as K. von Klitzing, G. Dorda & M. Pepper, New method for high-accuracy determination of the fine-structure constant based on quantised Hall resistance, *Physical Review Letters* **45**, pp. 494–497, 1980. Cited on page 791.

805 [3] The first paper was J. Bardeen, L.N. Cooper & J.R. Schrieffer, Microscopic Theory of Superconductivity, *Physical Review* **106**, pp. 162–164, 1957. The full, so-called BCS theory, is then given in the masterly paper J. Bardeen, L.N. Cooper & J.R. Schrieffer, Theory of Superconductivity, *Physical Review* **108**, pp. 1175–1204, 1957. Cited on page 789.

806 H.B.G. Casimir, On the attraction between two perfectly conducting bodies, *Proc. Kon. Ned. Acad. Wetensch.* **51**, pp. 793–795, 1948. No citations.

807 H.B.G. Casimir & D. Polder, The influence of retardation on the London-Van der Waals forces, *Physical Review* **73**, pp. 360–372, 1948. No citations.

808 S.L. Boersma, Aantrekkende kracht tussen schepen, *nederlands tidschrift voor natuuurkunde* **67**, pp. 244–246, Augustus 2001. See also his paper S.L. Boersma,A maritime analogy of the Casimir effect, *American Journal of Physics* **64**, pp. 539–541, 1996. Cited on page 796.

809 A. Larraza & B. Denardo, An acoustic Casimir effect, *Physics Letters A* **248**, pp. 151–155, 1998. See also A. Larraza, A demonstration apparatus for an acoustic analog of the Casimir effect, *American Journal of Physics* **67**, pp. 1028–1030, 1999. Cited on page 797.

810 M.J. Sparnaay, *Nature* **180**, p. 334, 1957. M.J. Sparnaay, Measurements of attractive forces between flat plates, *Physica (Utrecht)* **24**, pp. 751–764, 1958. Cited on page 798.

811 S.K. Lamoreaux, Demonstration of the Casimir force in the 0.6 to 6 μm range, *Physical Review Letters* **78**, pp. 5–8, 1997. U. Mohideen & A. Roy, Precision measurement of the Casimir force from 0.1 to 0.9 µm, *Physical Review Letters* **81**, pp. 4549–4552, 1998. Cited on page 798.

812 T.H. Boyer, Van der Waals forces and zero-point energy for dielectric and permeable materials, *Physical Review A* **9**, pp. 2078–2084, 1974. This was taken up again in O. Kenneth, I. Klich, A. Mann & M. Revzen, Repulsive Casimir forces, *Physical Review Letters* **89**, p. 033001, 2002. However, none of these effects has been verified yet.Cited on page 798.

813 The effect was discovered by K. Scharnhorst, On propagation of light in the vacuum between plates, *Physics Letters B* **236**, pp. 354–359, 1990. See also G. Barton, Faster-than-c light between parallel mirrors: the Scharnhorst effect rederived, *Physics Letters B* **237**, pp. 559–562, 1990, and P.W. Milonni & K. Svozil, Impossibility of measuring faster-than-c signalling by the Scharnhorst effect, *Physics Letters B* **248**, pp. 437–438, 1990. The latter corrects an error in the first paper. Cited on page 798.

814 D.L. Burke & al., Positron production in multiphoton light-by-light scattering, *Physical Review Letters* **79**, pp. 1626–1629, 1997. A simple summary is given in Bertram Schwarzschild, Gamma rays create matter just by plowing into laser light, *Physics Today* **51**, pp. 17–18, February 1998. Cited on page 801.

815 M. Kardar & R. Golestanian, The 'friction' of the vacuum, and other fluctuation-induced forces, *Reviews of Modern Physics* **71**, pp. 1233–1245, 1999. Cited on page 801.

816 G. Gour & L. Sriramkumar, Will small particles exhibit Brownian motion in the quantum vacuum?, http://www.arxiv.org/abs/quant/phys/9808032 Cited on page 801.

817 P.A.M. Dirac, ..., 1985. This article transcribes a speech by Paul Dirac just before his death. Disturbed by the infinities of quantum electrodynamics, he rejects as wrong the theory which he himself found, even though it correctly describes all experiments. The humility of a great man who is so dismissive of his main and great achievement was impressive when hearing him, and still is impressive when reading the speech. Cited on page 802.

818 See for example the overview by M. Born & T. Jüstel, Umweltfreundliche Lichtquellen, *Physik Journal* **2**, pp. 43–49, 2003. Cited on page 792.

819 J. Kasparian, Les lasers femtosecondes : applcations atmosphériques, *La Recherche* pp. RE14 1–7, February 2004/ See also the http://www.teramobile.org website. Cited on page 794.

820 Misleading statements are given in the introduction and in the conclusion of the review by H. Paul, Interference between independent photons, Review of Modern Physics, volume 58, pages 209-231; however, in the bulk of the article the author in practice retracts the statement, e.g. on page 221. Cited on page 794.

821 H. Euler & B. Kockel, *Naturwissenschaften* **23**, p. 246, 1935, H. Euler, *Annalen der Physik* **26**, p. 398, 1936, W. Heisenberg & H. Euler, Folgerung aus der Diracschen Theorie des Electrons, *Zeitschrift für Physik* **98**, pp. 714–722, 1936. Cited on page 802.

822 Thomas Stoehlker & al., *Physical Review Letters* 9 October 2000. Cited on page 803.

823 M. Roberts & al., Observation of an electric octupole transition in a single ion, *Physical Review Letters* **78**, pp. 1876–1879, 1997. A lifetime of 3700 days was determined. Cited on page 803.

824 K. An, J.J. Childs, R.R. Desari & M.S. Field, Microlaser: a laser with one atom in an optical resonator, *Physical Review Letters* **73**, pp. 3375–3378, 19 December 1994. Cited on page 794.

825 See the website by Tom Hales on http://www.math.lsa.umich.edu/~hales/countdown. An earlier book on the issue is the delightful text by Max Leppmaier, *Kugelpackungen von Kepler bis heute*, Vieweg Verlag, 1997. Cited on page 803.

826 For a short overview, also covering binary systems, see C. Bechinger, H.-H. von Grünberg & P. Leiderer, Entropische Kräfte - warum sich repulsiv wechselwirkende Teilchen anziehen können, *Physikalische Blätter* **55**, pp. 53–56, 1999. Cited on page 804.

827 Hans–Werner Fink, Direct observation of three-body interactions in adsorbed layers: Re on W(111), *Physical Review Letters* **52**, pp. 1532–1534, 1984. Cited on page 804.

828 See for example the textbook by A. Pimpinelli & J. Villain, *Physics of Crystal Growth*, Cambridge University Press, 1998. Cited on page 804.

829 Y. Furukawa, Faszination der Schneekristalle - wie ihre bezaubernden Formen entstehen, *Chemie in unserer Zeit* **31**, pp. 58–65, 1997. His http://www.lowtem.hokudai.ac.jp/~frkw/index_e.html web site gives more information. Cited on page 804.

830 P.-G. de Gennes, Soft matter, *Reviews of Modern Physics* **63**, p. 645, 1992. Cited on page 805.

831 The fluid helium based quantum gyroscope was first built by the Czech group ... and the Berkeley group, *Nature* 10. April 1997. Cited on page 805.

832 K. Schwab & al., *Nature* **404**, p. 974, 2000. For the optical analog, see reference Ref. 489. Cited on page 805.

833 J.P. Briand & al., Production of hollow atoms by the excitation of highly charged ions in interaction with a metallic surface, *Physical Review Letters* **65**, pp. 159–162, 1990. See also C. Rhodes, Hollow atoms, ..., and ...*Physikalische Blätter* Oktober 1996. Cited on page 805.

834 The present knowledge on how humans move is summarised by ... Cited on page 808.

835 See e.g. D.A. Bryant, Electron acceleration in the aurora, *Contemporary Physics* **35**, pp. 165–179, 1994. Cited on page 808.

836 The discharges above clouds are described e.g. by ... Cited on page 808.

837 See e.g. L.O'C. Drury, Acceleration of cosmic rays, *Contemporary Physics* **35**, pp. 231–242, 1994. Cited on page 808.

838 About Coulomb explosions, see ... Cited on page 808.

839 See Dirk Kreimer, New mathematical structures in renormalizable quantum field theories, http://www.arxiv.org/abs/hep-th/0211136 or *Annals of Physics* **303**, pp. 179–202, 2003, and the erratum ibid., **305**, p. 79, 2003. Cited on page 808.

840 R. Colella, A.W. Overhauser & S.A. Werner, Observation of gravitationally induced quantum mechanics, *Physical Review Letters* **34**, pp. 1472–1475, 1975. This experiment is usually called the cow experiment. Cited on page 809.

841 G.C. Ghirardi, A. Rimini & T. Weber, Unified dynamics for microscopic and macroscopic systems, *Physical Review D* **34**, pp. 470–491, 1986. Cited on page 810.

842 Ch.J. Bordé & C. Lämmerzahl, Atominterferometrie und Gravitation, *Physikalische Blätter* **52**, pp. 238–240, 1996. See also A. Peters, K.Y. Chung & S. Chu, Observation of gravitational acceleration by dropping atoms, *Nature* **400**, pp. 849–852, 26 August 1999. Cited on page 810.

843 V.V. Nesvizhevsky & al., Quantum states of neutrons in the Earth's gravitational field, *Nature* **415**, p. 297, 17 January 2002. Cited on page 810.

844 J.D. Bekenstein, Black holes and entropy, *Physical Review D* **7**, pp. 2333–2346, 1973. Cited on page 811.

845 R. Bousso, The holographic principle, *Review of Modern Physics* **74**, pp. 825–874, 2002, also available as http://www.arxiv.org/abs/hep-th/0203101. The paper is an excellent review; however, it has some argumentation errors, as explained on page 812. Cited on page 812.

846 This is clearly argued by S. Carlip, Black hole entropy from horizon conformal field theory, *Nuclear Physics B* Proceedings Supplement **88**, pp. 10–16, 2000. Cited on page 812.

847 S.A. Fulling, Nonuniqueness of canonical field quantization in Riemannian space-time, *Physical Review D* **7**, pp. 2850–2862, 1973, P. Davies/ Scalar particle production in Schwarzchild and Rindler metrics, *Journal of Physics A: General Physics* **8**, pp. 609–616, 1975, W.G. Unruh, Notes on black hole evaporation, *Physical Review D* **14**, pp. 870–892, 1976. Cited on page 812.

848 About the possibility to measure Fulling–Davies–Unruh radiation, see for example the paper by H. Rosu, On the estimates to measure Hawking effect und Unruh effect in the laboratory, *International Journal of Modern Physics D3* p. 545, 1994. http://www.arxiv.org/abs/gr-qc/9605032 or the paper P. Chen & T. Tajima, Testing Unruh radiation with ultraintense lasers, *Physical Review Letters* **83**, pp. 256–259, 1999. Cited on page 812.

849 W.G. Unruh & R.M. Wald, Acceleration radiation and the generalised second law of thermodynamics, *Physical Review D* **25**, pp. 942–958 , 1982. Cited on page 813.

850 R.M. Wald, The thermodynamics of black holes, *Living Reviews of Relativity* 2001, www-livingreviews.org/lrr-2001-6. Cited on page 814.

851 For a delightful popular account from a black hole expert, see IGOR NOVIKOV, *Black Holes and the Universe*, Cambridge University Press 1990. Cited on pages 815 and 818.

852 The original paper is W.G. UNRUH, Experimental black hole evaporation?, *Physical Review Letters* **46**, pp. 1351–1353, 1981. A good explanation with good literature overview is the one by MATT VISSER, Acoustical black holes: horizons, ergospheres and Hawking radiation, http://www.arxiv.org/abs/gr-qc/9712010. Cited on page 815.

853 G.W. GIBBONS & S.W. HAWKING, Action integrals and partition functions in quantum gravity, *Physical Review D* **15**, pp. 2752–2756, 1977. No citations.

854 Optical black holes ... Cited on page 815.

855 T. DAMOUR & R. RUFFINI, Quantum electrodynamical effects in Kerr–Newman geometries, *Physical Review Letters* **35**, pp. 463–466, 1975. Cited on page 815.

856 These were the Vela satellites; their existence and results were announced officially only in 1974, even though they were working already for many years. Cited on page 816.

857 The best general introduction into the topic of gamma ray bursts is S. KLOSE, J. GREINER & D. HARTMANN, Kosmische Gammastrahlenausbrüche – Beobachtungen und Modelle, Teil I und II, *Sterne und Weltraum* March and April 2001. Cited on page 816.

858 When the gamma ray burst encounters the matter around the black hole, it is broadened. The larger the amount of matter, the broader the pulse is. See G. PREPARATA, R. RUFFINI & S.-S. XUE, The dyadosphere of black holes and gamma-ray bursts, *Astronomy and Astrophysics* **338**, pp. L87–L90, 1998, R. RUFFINI, J.D. SALMONSON, J.R. WILSON & S.-S. XUE, On the pair electromagnetic pulse of a black hole with electromagnetic structure, *Astronomy and Astrophysics* **350**, pp. 334–343, 1999, R. RUFFINI, J.D. SALMONSON, J.R. WILSON & S.-S. XUE, On the pair electromagnetic pulse from an electromagnetic black hole surrounded by a baryonic remnant, *Astronomy and Astrophysics* **359**, pp. 855–864, 2000, and C.L. BIANCO, R. RUFFINI & S.-S. XUE, The elementary spike produced by a pure e^+e^- pair-electromagnetic pulse from a black hole: the PEM pulse, *Astronomy and Astrophysics* **368**, pp. 377–390, 2001. For a very personal account by Ruffini on his involvement in gamma ray bursts, see his paper Black hole formation and gamma ray bursts, http://www.arxiv.org/abs/astro-ph/0001425. Cited on page 816.

859 About the membrane despription of black holes, see ... Cited on page 818.

860 http://www.arxiv.org/abs/astro-ph/0301262, http://www.arxiv.org/abs/astro-ph/0303539, D.W. FOX & al., Early optical emission from the γ-ray burst of 4 October 2002, *Nature* **422**, pp. 284–286, 2003. Cited on page 816.

861 Negative heat capacity has also been found in atom clusters and in nuclei. See e.g. M. SCHMIDT & al., Negative heat capacity for a cluster of 147 sodium atoms, *Physical Review Letters* **86**, pp. 1191–1194, 2001. Cited on page 818.

862 H.-P. NOLLERT, Quasinormal modes: the characteristic 'sound' of black holes and neutron stars, *Classical and Quantum Gravity* **16**, pp. R159–R216, 1999. Cited on page 818.

863 He wrote a series of papers on the topic; a beautiful summary is DON N. PAGE, How fast does a black hole radiate information?, *International Journal of Modern Physics* **3**, pp. 93–106, 1994, based on his earlier papers, such as Information in black hole radiation, *Physical Review Letters* **71**, pp. 3743–3746, 1993, and Average entropy of a subsystem, *Physical Review Letters* **71**, pp. 1291–1294, 1993. The entropy formula of the latter paper, used above, was proven by S.K. FOONG & S. KANNO, Proof of Page's conjecture on the average entropy of a subsystem, *Physical Review Letters* **72**, pp. 1148–1151, 1994. See also his electronic preprint at http://www.arxiv.org/abs/hep-th/9305040. Cited on page 819.

864 See DON N. PAGE, ..., Cited on page 820.

865 The author speculates that this version of the coincidences could be original; he has not found it in the literature. Cited on page 810.

866 STEVEN WEINBERG, *Gravitation and Cosmology*, (Wiley, 1972). See equation 16.4.3 on page 619 and also page 620. Cited on page 810.

867 It could be that knot theory provides a relation between local knot invariant and a global one. Cited on page 811.

868 R. LAFRANCE & R.C. MYERS, Gravity's rainbow: limits for the applicability of the equivalence principle, *Physical Review D* **51**, pp. 2584–2590, 1995, http://www.arxiv.org/abs/hep-th/9411018. Cited on page 820.

869 M.YU. KUCHIEV & V.V. FLAMBAUM, Scattering of scalar particles by a black hole, http://www.arxiv.org/abs/gr-qc/0312065. See also M.YU. KUCHIEV & V.V. FLAMBAUM, Reflection on event horizon and escape of particles from confinement inside black holes, http://www.arxiv.org/abs/gr-qc/0407077. Cited on page 820.

870 E. JOOS, Why do we observe a classical space-time?, *Physics Letters A* **116**, pp. 6–8, 1986. Cited on page 821.

871 This point was made repeatedly by STEVEN WEINBERG, namely in Derivation of gauge invariance and the equivalence principle from Lorentz invariance of the S-matrix , *Physics Letters* **9**, pp. 357–359, 1964, in Photons and gravitons in S-matrix theory: derivation of charge conservation and equality of gravitational and inertial mass, *Physical Review B* **135**, pp. 1049–1056, 1964, and in Photons and gravitons in perturbation theory: derivation of Maxwell's and Einstein's equations, *Physical Review B* **138**, pp. 988–1002, 1965. Cited on page 822.

872 L.H. FORD & T.A. ROMAN, Quantum field theory constrains traversable wormhole geometries, http://www.arxiv.org/abs/gr-qc/9510071 or *Physical Review D* **53**, pp. 5496–5507, 1996. Cited on page 823.

873 M.J. PFENNING & L.H. FORD, The unphysical nature of 'warp drive', *Classical and Quantum Gravity* **14**, pp. 1743–1751, 1997. Cited on page 823.

874 B.S. KAY, M. RADZIKOWSKI & R.M. WALD, Quantum field theory on spacetimes with a compactly generated Cauchy horizon, http://www.arxiv.org/abs/gr-qc/9603012 or *Communications in Mathematical Physics 183* pp. 533–556, 1997. Cited on page 823.

875 An overview on squeezed light is given in ... No citations.

876 An introduction into the history of the concept of precision is given by N.W. WISE, The values of precision, Princeton University Press, 1994. No citations.

877 E.H. KENNARD, Zur Quantenmechanik einfacher Bewegungstypen, *Zeitschrift für Physik* **44**, pp. 326–352, 1927. The priority of the result is often wrongly attributed to the much later paper by H.P. ROBERTSON, The uncertainty principle, *Physical Review* **34**, pp. 163–164, 1929. The complete result is by ERWIN SCHRÖDINGER, Zum Heisenbergschem Unschärfeprinzip, *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse* Berlin, pp. 296–303, 1930. No citations.

878 W. HEISENBERG, Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik, *Zeitschrift für Physik* **43**, pp. 172–198, 1927. No citations.

879 M.G. RAYMER, Uncertainty principle for joint measurement of noncommuting variables, *American Journal of Physics* **62**, pp. 986–993, 1994. No citations.

880 The question about the origin of blue colour in flowers has been discussed already in the beginning of the twentieth century. A short summary is given by K. YOSHIDA & al., Cause of blue petal colour, *Nature* **373**, p. 291, 26 January 1995. No citations.

881 The main reference on nondemolition measurements is the text by C.M. Caves & P.D. Drumond, Quantum limits on bosonic communication rates, *Reviews of Modern Physics* **66**, pp. 481–537, 1994. No citations.

882 Mark P. Silverman, *And Yet it Moves: Strange Systems and Subtle Questions in Physics*, Cambridge University Press, 1993. It is a treasure chest for anybody interested in the details of modern physics.

The first of the two other books by Mark P. Silverman, *More than one Mystery*, Springer 1995, talks about the Aharonov-Bohm effect, about the EPR paradox and about the Hanbury–Brown–Twiss experiments; his book *Waves and Grains: Reflections on Light and Learning*, Princeton University Press, 1998, completes his exposition of interesting thoughts and insights about physics. No citations.

883 N. Ashcroft & D. Mermin, *Solid State Physics*, Saunders, 1976, is probably the best as well as the most original and entertaining introduction to the topic available in English language. No citations.

884 The famous quantum theory book is P.A.M. Dirac, Quantum Mechanics, 4th revised edition, Oxford University Press, 1967. No citations.

Chapter VIII

Inside the Nucleus

29. The structure of the nucleus – the densest clouds

Nuclear physics was born in 1896 in France, but is now a small activity. It produced not more than one daughter, experimental high energy physics, which was born around 1930. But since 1985, also these activities are in strong decline. Despite the shortness, the family history is impressive; the two fields uncovered why stars shine, how powerful bombs work, how cosmic evolution produced the atoms we are made of and how medical doctors can dramatically improve their healing rate. Ref. 885

> Nuclear physics is just low-density astrophysics.

A physical wonder: magnetic resonance imaging

Arguably, the most spectacular tool that physical research has produced in the twentieth century was magnetic resonance imaging, or MRI for short. This technique allows to image human bodies with a high resolution and with (almost) no damage, in strong contrast to X-ray imaging. Though the machines are still expensive – costing 400 000 Euro and more – there is hope that they will become cheaper in the future. Such a machine consists essentially of a large magnetic coil, a radio transmitter and a computer. Some results of putting part of a person into the coil are shown in Figure 332.

In these machines, a radio transmitter emits radio waves that are absorbed because hydrogen nuclei are small spinning magnets. The magnets can be parallel or antiparallel to the magnetic field produced by the coil. The transition energy E can be absorbed from a radio wave whose frequency ω is tuned to the magnetic field B. The energy absorbed by a single hydrogen nucleus is given by

$$E = \hbar\omega = \hbar\gamma B \qquad (607)$$

The material constant $\gamma/2\pi$ has a value of 42.6 MHz/T for hydrogen nuclei; it results from the non-vanishing spin of the proton. This is a quantum effect, as stressed by the appearance of the quantum of action $\hbar$. Using some cleverly applied magnetic fields, typically with a strength between 0.3 and 1.5 T, the machines are able to measure the absorption for each volume element separately. Interestingly, the precise absorption level depends on the chemical compound the nucleus is built into. Thus the absorption value will depend on the chemical environment. When the intensity of the absorption is plotted as grey scale, an image is formed that retraces the different chemical composition. Two examples are shown in Figure 332. Using additional tricks, modern machines can picture blood flow

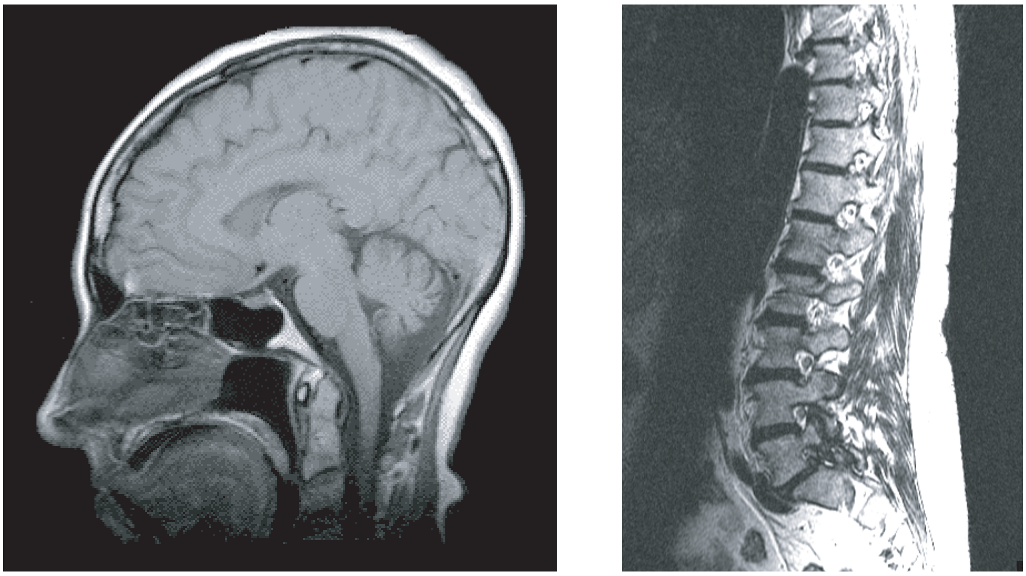


Figure 332 Sagittal images of the head and the spine – used with permission from Joseph P. Hornak, *The Basics of MRI*, http://www.cis.rit.edu/htbooks/mri, Copyright 2003

in the heart or air flow in lungs; they can even make movies of the heart beat. Other techniques show how the location of sugar metabolism in the brain depends on what you are thinking about.* In fact, also what you are thinking about all the time has been imaged: Ref. 886 the first image of people making love has been taken by Willibrord Weijmar Schultz and his group in 1999. It is shown in Figure 333.

Each magnetic resonance image thus proves that atoms have spinning nuclei. Like for any other object, nuclei have size, colour, composition and interactions that ask to be explored.

The size of nuclei

The magnetic resonance signal shows that hydrogen nuclei are quite sensitive to magnetic fields. The g-factor of protons, defined using the magnetic moment μ, their mass and Page 710 charge as $g = \mu 4m/e\hbar$, is about 5.6. Using expression (534) that relates the g-factor and the radius of a composite object, we deduce that the radius of the proton is about 0.9 fm; this value is confirmed by experiment. Protons are thus much smaller than hydrogen atoms, the smallest of atoms, whose radius is about 30 pm. In turn, the proton is the smallest of all nuclei; the largest nuclei have radii 7 times the proton value.

The small size of nuclei is no news. It is known since the beginning of the twentieth century. The story starts on the first of March in 1896, when Henri Becquerel* dis-

* The website http://www.cis.rit.edu/htbooks/mri by Joseph P. Hornak gives an excellent introduction to magnetic resonance imaging, both in English and Russian, including the physical basis, the working of the machines, and numerous beautiful pictures. The method of studying nuclei by putting them at the same time into magnetic and radio fields is also called *nuclear magnetic resonance*.

* Henri Becquerel (b. 1852 Paris, d. 1908 Le Croisic), important French physicist; his primary topic was

covered a puzzling phenomenon: minerals of uranium potassium sulphate blacken photographic plates. Becquerel had heard that the material is strongly fluorescent; he conjectured that fluorescence might have some connection to the X-rays discovered by Conrad Röngten the year before. His conjecture was wrong; nevertheless it led him to an important new discovery. Investigating the reason for the effect of uranium on photographic plates, Becquerel found that these minerals emit an undiscovered type of radiation, different from anything known at that time; in addition, the radiation is emitted by any substance containing uranium. In 1898, Bémont named the property of these minerals *radioactivity*.

Radioactive rays are also emitted from many elements other than uranium. The radiation can be 'seen': it can be detected by the tiny flashes of light that are emitted when the rays hit a scintillation screen. The light flashes are tiny even at a distance of several meter from the source; thus the rays must be emitted from point-like sources. Radioactivity has to be emitted from single atoms. Thus radioactivity confirmed unambiguously that atoms do exist. In fact, radioactivity even allows to count them, as we will find out shortly.

Henri Becquerel

The intensity of radioactivity cannot be influenced by magnetic or electric fields; it does not depend on temperature or light irradiation. In short, radioactivity does not depend on electromagnetism and is not related to it. Also the high energy of the emitted radiation cannot be explained by electromagnetic effects. Radioactivity must thus be due to another, new type of force. In fact, it took 30 years and a dozen of Nobel prizes to fully understand the details. It turns out that several types of radioactivity exist; the types behave differently when they fly through a magnetic field or when they encounter matter. They are listed in Table 63. All have been studied in great detail, with the aim to understand the nature of the emitted entity and its interaction with matter.

In 1909, radioactivity inspired the 37 year old physicist Ernest Rutherford,* who had won the Nobel prize just the year before, to another of his brilliant experiments. He asked his collaborator Hans Geiger to take an emitter of alpha radiation – a type of radioactivity which Rutherford had identified and named 10 years earlier – and to point the radiation at a thin metal foil. The quest was to find out where the alpha rays would end up. The research group followed the path of the particles by using scintillation screens; later on they used

Marie Curie

the study of radioactivity. He was the thesis adviser of Marie Curie, the wife of Pierre Cu[illegible]al to bringing her to fame. The SI unit for radioactivity is named after him. For his discovery of radioactivity he received the 1903 Nobel prize for physics; he shared it with the Curies.

* Ernest Rutherford (1871–1937), important New Zealand physicist. He emigrated to Britain and became professor at the University of Manchester. He coined the terms alpha particle, beta particle, proton and neutron. A gifted experimentalist, he discovered that radioactivity transmutes the elements, explained the nature of alpha rays, discovered the nucleus, measured its size and performed the first nuclear reactions. Ironically, in 1908 he received the Nobel price for chemistry, much to the amusement of himself and of the world-wide physics community; this was necessary as it was impossible to give enough physics prizes to the numerous discoverers of the time. He founded a successful research school of nuclear physics and many famous physicists spent some time at his institute. Ever an experimentalist, Rutherford deeply disliked quantum theory, even though it was and is the only possible explanation for his discoveries.

Figure 333 The origin of human life (© Willibrord Weijmar Schultz)

an invention by Charles Wilson: the cloud chamber. A *cloud chamber*, like its successor, the bubble chamber, produces white traces along the path of charged particles; the mechanism is the same as the one than leads to the white lines in the sky when an aeroplane flies by.

The radiation detectors gave a strange result: most alpha particles pass through the metal foil undisturbed, whereas a few are reflected. In addition, those few which are reflected are not reflected by the surface, but in the inside of the foil. (Can you imagine how they showed this?) Rutherford deduced from this scattering experiment that first of all, atoms are mainly transparent. Only transparency explains why most alpha particles pass the foil without disturbance, even though it was over 2000 atoms thick. But some particles were scattered by large angles or even reflected. Rutherford showed that the reflections must be due to a single scattering point. By counting the particles that were reflected (about 1 in 20000 for his 0.4 μm gold foil), Rutherford was also able to deduce the size of the reflecting entity and to estimate its mass. He found that it contains almost all of the mass of the atom in a diameter of around 1 fm. He thus named it the *nucleus*. Using the knowledge that atoms contain electrons, Rutherford then deduced from this experiment that atoms consist of an electron cloud that determines the size of atoms – of the order of 0.1 nm – and of a tiny but heavy nucleus at the centre. If an atom had the size of a basketball, its nucleus would have the size of a dust particle, yet contain 99.9 % of the basketball's mass. Atoms resemble thus candy floss around a heavy dust particle. Even though the candy floss – the electron cloud – around the nucleus is extremely thin and light, it is strong enough to avoid that two atoms interpenetrate; thus it keeps the neighbouring nuclei at constant distance. For the tiny and massive alpha however, particles the candy floss is essentially empty space, so that they simply fly through the electron clouds until they exit on the other side or hit a nucleus.

Challenge 1336 n

Table 63 The main types of radioactivity and rays emitted by matter

Type	Particle	Example	Range	Danger	Shield	Use
α rays 3 to 10 MeV	helium nuclei	^{235}U, ^{238}U, ^{238}Pu, ^{238}Pu, ^{241}Am	a few cm in air	when eaten, inhaled, touched	any material, e.g. paper	thickness measurement
β rays 0 to 5 MeV	electrons and	^{14}C, ^{40}K, ^{3}H, ^{101}Tc	<1 mm in metal	serious	metals	cancer treatment
	antineutrinos		light years	none	none	research
β^+ rays	positrons and	^{40}K, ^{11}C, ^{11}C, ^{13}N, ^{15}O	less than β	medium	any material	tomography
	neutrinos		light years	none	none	research
γ rays	high energy photons	^{110}Ag	several m in air	high	thick lead	preservation of herbs, disinfection
n reactions *c.* 1 MeV	neutrons	^{252}Cf, Po-Li $(\alpha$,n), ^{38}Cl-Be (γ,n)	...	high	0.3 m of paraffin	...
n emission typ. 40 MeV	neutrons	^{9}He, ^{24}N	...	high	0.3 m of paraffin	...
p emission typ. 20 MeV	protons	^{5}Be, ^{161}Re	like α rays	small	solids	
spontaneous fission typ. 100 MeV	nuclei	^{232}Cm, ^{263}Rf	like α rays	small	solids	detection of new elements

The density of the nucleus is impressive: about $5.5 \cdot 10^{17}$ kg/m^3. At that density, the mass of the Earth would fit in a sphere of 137 m radius and a grain of sand would have a mass larger than the largest existing oil tanker. (True?) Now we know that oil tankers are complex structures. What then is the structure of a nucleus?

Challenge 1337 e

I now know how an atom looks like!

Ernest Rutherford

Nuclei are composed

The magnetic resonance images also show that nuclei are composed. Images can be taken also using heavier nuclei instead of hydrogen, such as certain fluorine or oxygen nuclei. The g-factors of these nuclei also depart from the value 2 characteristic of point particles;

Page 710 the more massive they are, the bigger the departure. Such objects have a finite size; indeed, the size of nuclei can be measured directly and confirm the values predicted by the g-factor. Both the values of the g-factor and the non-vanishing sizes show that nuclei are composed.

Interestingly, the idea that nuclei are composed is older than the concept of nucleus itself. Already in 1815, after the first mass measurements of atoms by John Dalton and others, researchers noted that the mass of the various chemical elements seem to be almost perfect multiples of the weight of the hydrogen atom. William Prout then formulated the hypothesis that all elements are composed of hydrogen. When the nucleus was discovered, knowing that it contains almost all mass of the atom, it was therefore first thought that all nuclei are made of hydrogen nuclei. Being at the origin of the list of constituents, the hydrogen nucleus was named *proton*, from the greek term for 'first' and reminding the name of Prout at the same time. Protons carry a positive unit of electric charge, just the opposite of that of electrons, but are almost 2000 times as heavy.

However, the charge and the mass numbers of the other nuclei do not match. On average, a nucleus that has n times the charge of a proton, has a mass that is about 2.6 n times than of the proton. Additional experiments then confirmed an idea formulated by Werner Heisenberg: all nuclei heavier than hydrogen nuclei are made of positively charged protons and neutral *neutrons*. Neutrons are particles a tiny bit more massive than protons (the difference is less than a part in 700), but without any electrical charge. Since the mass is almost the same, the mass of nuclei – and thus that of atoms – is still an (almost perfect) integer multiple of the proton mass. But since neutrons are neutral, the mass and the charge number of nuclei differ. Being neutral, neutrons do not leave tracks in clouds chambers and are more difficult to detect. For this reason, they were discovered much later than other subatomic particles.

Today it is possible to keep single neutrons suspended between suitably shaped coils, with the aid of teflon 'windows'. Such traps were proposed in 1951 by Wolfgang Paul. They work because neutrons, though they have no charge, do have a small magnetic moment. (By the way, this implies that neutrons are composed of charged particles.) With a suitable arrangement of magnetic fields, neutrons can be kept in place, in other words, they can be levitated. Obviously, a trap only makes sense if the trapped particle can be observed. In case of neutrons, this is achieved by the radio waves absorbed when the magnetic moment switches direction with respect to an applied magnetic field. The result of these experiments is simple: the lifetime of free neutrons is around 888(1) s. Nevertheless, inside most nuclei we are made of, neutrons do not decay, as the result does not lead to a state of lower energy. (Why not?) Challenge 1338 n

Magnetic resonance images also show that some elements have different types of atoms. These elements have atoms that with the same number of protons, but with different numbers of neutrons. One says that these elements have several isotopes.* This also explains why some elements radiate with a mixture of different decay times. Though chemically they are (almost) indistinguishable, isotopes can differ strongly in their nuclear properties. Some elements, such as tin, caesium, or polonium, have over thirty iso-

* The name is derived from the Greek words for 'same' and 'spot', as the atoms are on the same spot in the periodic table of the elements.

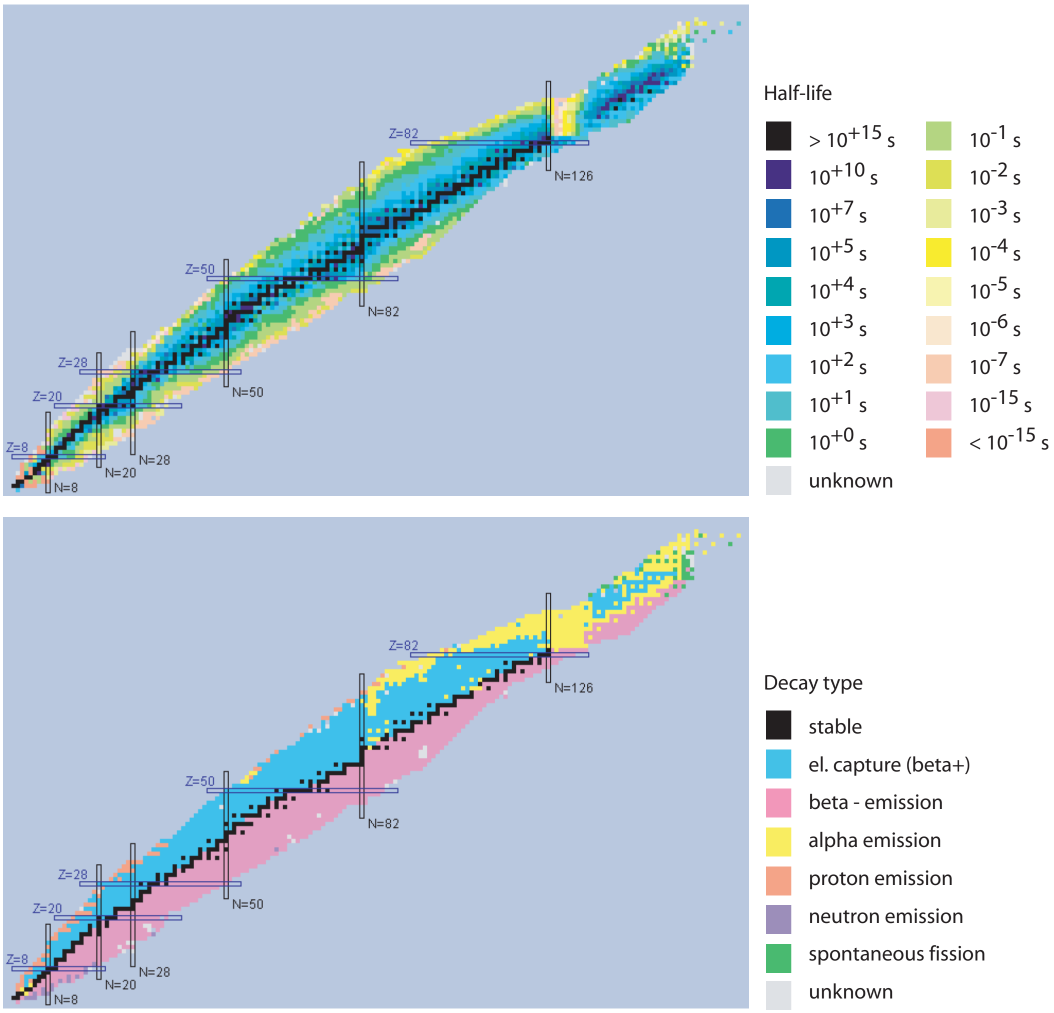


Figure 334 All known nuclides with their lifetimes and main decay modes (data from http://www.nndc.bnl.gov/nudat2)

topes each. Together, the about 100 known elements have over 2000 *nuclides*.*

The motion of protons and neutrons inside nuclei allows to understand the spin and the magnetic moment of nuclei. Since nuclei are so extremely dense despite containing numerous positively charged protons, there must be a force that keeps everything together against the electrostatic repulsion. We saw that the force is not influenced by electromagnetic or gravitational fields; it must be something different. The force must be short range; otherwise nuclei would not decay by emitting high energy alpha rays. The new force is called the strong nuclear interaction. We shall study it in detail shortly.

* Nuclides is the standard expression for a nucleus with a given number of neutrons and protons.

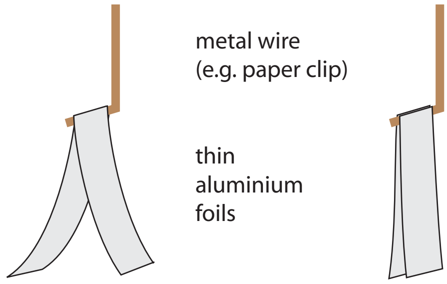


Figure 335 An electroscope (or electrometer) in charged (left) and charged state (right)

Nuclei can move alone – cosmic rays

In everyday life, nuclei are mostly found inside atoms. But in some situations, they move all by themselves. The first to discover an example was Rutherford, who had shown that alpha particles are helium nuclei. Like all nuclei, alpha particles are small, so that they are quite useful as projectiles.

Then, in 1912, Viktor Heß* made a completely unexpected discovery. Heß was intrigued by electroscopes (also called electrometers). These are the simplest possible detectors of electric charge. They mainly consist of two hanging, thin metal foils, such as two strips of aluminium foil taken from a chocolate bar. When the electroscope is charged, the strips repel each other and move apart, as shown in Figure 335. (You can build one easily yourself by covering an empty glass with some transparent cellophane foil and suspending a paper clip and the aluminium strips from the foil.) An electroscope thus measures electrical charge. Like many before him, Heß noted that even for a completely isolated electroscope, the charge disappears after a while. He asked: why? By careful study he eliminated one explanation after the other, he and others were left with only one possibility: that the discharge could be due to charged rays, such as those of the recently discovered radioactivity. He thus prepared a sensitive electrometer and took it with him on a balloon flight.

Challenge 1339 e

Viktor Heß

As expected, the balloon flight showed that the discharge effect diminished with height, due to the larger distance from the radioactive substances on the Earth's surface.

* Viktor Franz Heß, (1883–1964), Austrian nuclear physicist, received the Nobel prize for physics in 1936 for his discovery of cosmic radiation. Heß was one of the pioneers of research into radioactivity. Heß' discovery also explained why the atmosphere is always somewhat charged, a result important for the formation and behaviour of clouds. Twenty years after the discovery of cosmic radiation, in 1932 Carl Anderson discovered the first antiparticle, the positron, in cosmic radiation; in 1937 Seth Neddermeyer and Carl Anderson discovered the muon; in 1947 a team led by Cecil Powell discovered the pion; in 1951, the Λ^0 and the kaon K^0 are discovered. All discoveries used cosmic rays and most of these discoveries led to Nobel prizes.

Table 64 The main types of cosmic radiation

PARTICLE	ENERGY	ORIGIN	DETECTOR	SHIELD
At high altitude, the primary particles:				
Protons (90 %)	10^9 to 10^{22} eV	stars, supernovae, extragalactic, unknown	scintillator	in mines
Alpha rays (9 %)	...	...	...	...
Other nuclei, such as iron (1 %)	10^9 to 10^{19} eV	stars, novae	...	...
Neutrinos	MeV, GeV	Sun, stars	chlorine, gallium, water	none
Electrons (0.1 %)	10^6 to $> 10^{12}$ eV	supernova remnants		
Gammas (10^{-6})	1 eV to 50 TeV	stars, pulsars, galactic, extragalactic	semiconductor detectors	in mines
At sea level, secondary particles are produced in the atmosphere:				
Muons	3 GeV, $150/\mathrm{m}^2\mathrm{s}$	protons hit atmosphere, produce pions which decay into muons	drift chamber	15 m of water or 2.5 m of soil
Oxygen and other nuclei				
Positrons				
Neutrons	...			
Pions	...			
In addition, there are slowed down primary beam particles.				

But above about 1000 m of height, the discharge effect increased again, and the higher he flew, the stronger it became. Risking his health and life, he continued upwards to more than 5000 m; there the discharge effect was several times stronger than on the surface of the Earth. This result is exactly what is expected from a radiation coming from outer space and absorbed by the atmosphere. In one of his most important flights, performed during an (almost total) solar eclipse, Heß showed that most of the 'height radiation' did not come from the Sun, but from further away. He – and Millikan – thus called the radiation *cosmic rays*. During the last few centuries, many people have drunken from a glass and eaten chocolate; but only Heß combined these activities with such careful observation and deduction that he earned a Nobel prize.*

Today, the most impressive detectors for cosmic rays are Geiger–Müller counters and spark chambers. Both share the same idea; a high voltage is applied between two metal parts kept in a thin and suitably chosen gas (a wire and a cylindrical mesh for the Geiger-Müller counter, two plates or wire meshes in the spark chambers). When a high energy

* In fact, Hess gold foils in his electrometer.

ionizing particle crosses the counter, a spark is generated, which can either be observed through the generated spark (as you can do yourself in the entrance hall of the CERN main building), or detected by the sudden current flow. Historically, the current was first amplified and sent to a loudspeaker, so that the particles can be heard by a 'click' noise. With a Geiger counter, one cannot see atoms or particles, but one can hear them. Finally, ionized atoms could be counted. Finding the right gas mixture is tricky; it is the reason that the counter has a double name. One needs a gas that extinguishes the spark after a while, to make the detector ready for the next particle. Müller was Geiger's assistant; he made the best counters by adding the right mixture of alcohol to the gas in the chamber. Nasty rumours maintained that this was discovered when another assistant tried, without success, to build counters while Müller was absent. When Müller, supposedly a heavy drinker, came back, everything worked again. However, the story is apocryphal. Today, Geiger–Müller counters are used around the world to detect radioactivity; the smallest fit in mobile phones and inside wrist watches.

The particle energy in cosmic rays spans a large range between 10^3 eV and at least 10^{20} eV; the latter is the same energy as a tennis ball after serve. Understanding the origin of cosmic rays is a science by its own. Some are galactic in origin, some are extragalactic. For most energies, supernova remnants – pulsars and the like – seem the best candidates. However, the source of the highest energy particles is still unknown.

In other words, cosmic rays are probably the only type of radiation discovered without the help of shadows. But shadows have been found later on. In a beautiful experiment performed in 1994, the shadow thrown by the Moon on high energy cosmic rays (about 10 TeV) was studied. When the position of the shadow is compared with the actual position of the Moon, a shift is found. Due to the magnetic field of the Earth, the cosmic ray Moon shadow would be shifted westwards for protons and eastwards for antiprotons. The data are consistent with a ratio of antiprotons between 0 % and 30 %. By studying the shadow, the experiment thus showed that high energy cosmic rays are mainly positively charged and thus consist mainly of matter, and only in small part, if at all, of antimatter.

Ref. 902

Figure to be included

Figure 336 A Geiger–Müller counter

Detailed observations showed that cosmic rays arrive on the surface of the Earth as a mixture of many types of particles, as shown in Table 64. They arrive from outside the atmosphere as a mixture of which the largest fraction are protons, alpha particles, iron and other nuclei. Nuclei can thus travel alone over large distances. The number of charged cosmic rays depends on their energy. At the lowest energies, charged cosmic rays hit the human body many times a second. The measurements also show that the rays arrive in irregular groups, called *showers*.

The distribution of the incoming direction of cosmic rays shows that many rays must be extragalactic in origin. The typical nuclei of cosmic radiation are ejected from stars and accelerated by supernova explosions. When they arrive on Earth, they interact with the atmosphere before they reach the surface of the Earth. The detailed acceleration mechanisms are still a topic of research.

Figure 337 An aurora borealis produced by charged particles in the night sky

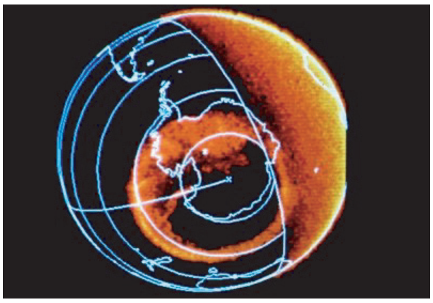
Cosmic rays have several effects on everyday life. Through the charges they produce in the atmosphere, they are probably responsible for the non-straight propagation of lightning. Cosmic rays are also important in the creation of rain drops and ice particles inside clouds, and thus indirectly in the charging of the clouds. Cosmic rays, together with ambient radioactivity, also start the Kelvin generator.

Page 480

If the Moon would not exist, we would die from cosmic rays. The Moon helps to give the Earth a high magnetic field via a dynamo effect, which then diverts most rays towards the magnetic poles. Also the upper atmosphere helps animal life to survive, by shielding life from the harmful effects of cosmic rays. Indeed, aeroplane pilots and airline employees have a strong radiation exposure that is not favourable to their health. Cosmic rays are one of several reasons that long space travel, such as a trip to mars, is not an option for humans. When cosmonauts get too much radiation exposure, the body weakens and eventually they die. Space heroes, including those of science fiction, would not survive much longer than two or three years.

Cosmic rays also produce beautifully coloured flashes inside the eyes of cosmonauts; they regularly enjoy these events in their trips. But cosmic rays are not only dangerous and beautiful. They are also useful. If cosmic rays would not exist at all, we would not exist either. Cosmic rays are responsible for mutations of life forms and thus are one of the causes of biological evolution. Today, this effect is even used artificially; putting cells into a radioactive environment yields new strains. Breeders regularly derive new mutants in this way.

Cosmic rays cannot be seen directly, but their cousins, the 'solar' rays, can. This is most spectacular when they arrive in high numbers. In such cases, the particles are inevitably deviated to the poles by the magnetic field of the Earth and form a so-called *aurora borealis* (at the North Pole) or an *aurora australis* (at the South pole). These slowly moving and variously coloured curtains of light belong to the most spectacular effects in the



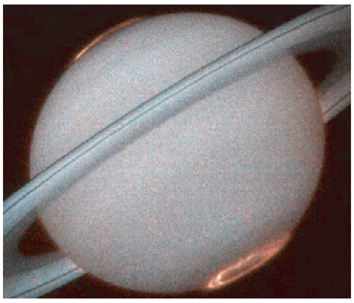


Figure 338 An aurora australis on Earth seen from space (in the X-ray domain) and one on Saturn

night sky. Visible light and X-rays are emitted at altitudes between 60 and 1000 km. Seen from space, the aurora curtains typically form a circle with a few thousand kilometres diameter around the magnetic poles.*

Cosmic rays are mainly free nuclei. With time, researchers found that nuclei appear without electron clouds also in other situations. In fact, the vast majority of nuclei in the universe have no electron clouds at all: in the inside of stars no nucleus is surrounded by bound electrons; similarly, a large part of intergalactic matter is made of protons. It is known today that most of the matter in the universe is found as protons or alpha particles inside stars and as thin gas between the galaxies. In other words, in contrast to what the Greeks said, matter is not usually made of atoms; it is mostly made of nuclei. Our everyday environment is an exception when seen on cosmic scales. In nature, atoms are rare.

By the way, nuclei are in no way forced to move; nuclei can also be stored with almost no motion. There are methods – now commonly used in research groups – to superpose electric and magnetic fields in such a way that a single nucleus can be kept floating in mid-air; we discussed this possibility in the section on levitation earlier on. Page 555

Nuclei decay

Not all nuclei are stable over time. The first measurement that provided a hint was the way radioactivity changes with time. The number N of atoms decreases with time. More precisely, radioactivity follows an exponential decay:

$$N(t) = N(0)e^{-t/\tau} \tag{608}$$

The parameter τ, the so-called *life time*, depends on the type of nucleus emitting the rays. It can vary from much less than a microsecond to millions of millions of years. The expression has been checked for as long as 34 multiples of the duration τ; its validity and precision is well-established by experiments. Radioactivity is the decay of unstable nuclei. Formula (608) is an approximation for large numbers of atoms, as it assumes that $N(t)$ is a continuous variable. Despite this approximation, deriving this expression from quantum theory is not a simple exercise, as we saw in the section on atomic physics. Though the Page 776

* In the solar system, aurorae due to core magnetic fields have been observed on Jupiter, Saturn, Uranus, Neptune, Earth, Io and Ganymede. Aurorae due to other mechanisms have been seen on Venus and Mars.

quantum Zeno effect can appear for small times t, for the case of radioactivity it has not been observed so far.

Most of all, the expression (608) allows to count the number of atoms in a given mass of material. Imagine to have measured the mass of radioactive material at the beginning of your experiment; you have chosen an element that has a lifetime of about a day. Then you put the material inside a scintillation box. After a few weeks the number of flashes has become so low that you can count them; using the formula you can then determine how many atoms have been in the mass to begin with. Radioactivity thus allows us to determine the number of atoms, and thus their size, in addition to the size of nuclei.

The decay (608) and the release of energy is typical of metastable systems. In 1903, Rutherford and Soddy discovered what the state of lower energy is for alpha and beta emitters. In these cases, radioactivity changes the emitting atom; it is a spontaneous transmutation of the atom. An atom emitting alpha or beta rays changes its chemical nature. Page 228 Radioactivity confirms what statistical mechanics of gases had concluded long time before: atoms have a structure that can change. In alpha decay, the radiating nucleus emits a (doubly charged) helium nucleus. The kinetic energy is typically a handful of MeV. After the emission, the nucleus has changed to one situated two places earlier in the periodic system of the elements.

In beta decay, a neutron transforms itself into a proton, emitting an electron and an antineutrino. Also beta decay changes the chemical nature of the atom, but to the place following the original atom in the periodic table of the elements. A variation is the beta+ decay, in which a proton changes into a neutron and emits a neutrino and a positron. We Page 869 will study these important decay processes below.

In gamma decay, the nucleus changes from an excited to a lower energy state by emitting a high energy photon. In this case, the chemical nature is not changed. Typical energies are in the MeV range. Due to the high energy, such rays ionize the material they encounter; since they are not charged, they are not well absorbed by matter and penetrate deep into materials. Gamma radiation is thus by far the most dangerous type of (outside) radioactivity.

By the way, in every human body about nine thousand radioactive decays take place every second, mainly 4.5 kBq (0.2 mSv/a) from ^{40}K and 4 kBq from ^{14}C (0.01 mSv/a). Challenge 1340 n Why is this not dangerous?

As a result of the chemical effects of radioactivity, the composition ratio of certain elements in minerals allows to determine the age of the mineral. For the first time it became possible to reliably date the age of rocks, to compare it with the age of meteorites and, when space travel became fashionable, with the age of the Moon. The result was beyond all estimates and expectations: the oldest rocks and the oldest meteorites studied independently using different dating methods, are 4570(10) million years old.

But if the Earth is that old, why did the Earth not cool down in its core in the meantime? The answer is radioactivity: the centre of the Earth is still hot because it contains radioactive potassium ^{40}K,* radioactive uranium ^{235}U and ^{238}U and radioactive thorium ^{232}Th. The radioactivity of these elements, and to minor degree a few others, keeps the Ref. 887 centre of the Earth glowing.

* The decay of potassium is the origin for the 1 % of argon found in the Earth's atmosphere.

Using radioactive dating, geologists determined the age of mountains, the age of sediments and the age of the continents. They calculated the time the Earth needed to cool down, the time that continents moved apart and the time that mountains formed when the continents collided. The times are consistent with the relative time scale that geologists had defined independently. All fell into place. Using the *radiocarbon method*, historians determined the age of civilizations and the age of human artefacts. Many beliefs were shattered. In some communities the shock is still not over, even though over hundred years have passed since these results became known.

What distinguishes those atoms that decay from those which do not? An exponential decay law implies that the probability of decay is independent of the age of the atom. Age or time plays no role. We also know from thermodynamics, that all atoms are exactly identical. So how is the decaying atom singled out? It took around 40 years to discover that decays are triggered by the statistical fluctuations of the vacuum, as described by quantum theory. Indeed, radioactivity is one of the clearest observations that classical physics is not sufficient to describe nature. Radioactivity, like all decays, is a pure quantum effect. Only a finite quantum of action makes it possible that a system remains unchanged until it suddenly decays. Indeed, in 1928 George Gamow explained alpha decay with the tunnelling effect. The tunnelling effect explains the relation between the lifetime and the range of the rays, as well as the measured variation of lifetimes – between 10 ns and 10^{17} years – as the consequence of the varying potentials to be overcome.

Page 720

By the way, massless particles cannot decay. There is a simple reason for it: massless particles do not experience time, as their paths are null. A particle that does not experience time cannot have a half-life. (Can you find another argument?)

Challenge 1341 ny

Nuclei can form composites

Nuclei are highly unstable when they contain more than about 280 nucleons. Higher mass values inevitably decay into smaller fragments. But when the mass is above 10^{57} nucleons, nuclear composites are stable again: such systems are then called neutron stars. This is the most extreme example of pure nuclear matter found in nature. Neutron stars are left overs of (type II) supernova explosions. They do not run any fusion reactions any more, as other stars do; in first approximation they are simply a large nucleus.

Neutron stars are made of degenerate matter. Their density of 10^{18} kg/m^3 is a few times that of a nucleus, as gravity compresses the star. This density value means that tea spoon of such a star has a mass of several 100 million tons. Neutron stars are about 10 km in diameter. They are never much smaller, as such stars are unstable. They are never much larger, because more massive neutron stars turn into black holes.

Nuclei have colours and shapes

In everyday life, the colour of objects is determined by the wavelength of light that is least absorbed, or if they shine, by the wavelength that is emitted. Also nuclei can absorb photons of suitably tuned energies and get into an excited state. In this case, the photon energy is converted into a higher energy of one or several of the nucleons whirling around inside the nucleus. Many radioactive nuclei also emit high energy photons, which then are called gamma rays, in the range of 1 keV (or 0.2 fJ) to about 20 MeV (or 3.3 pJ). The

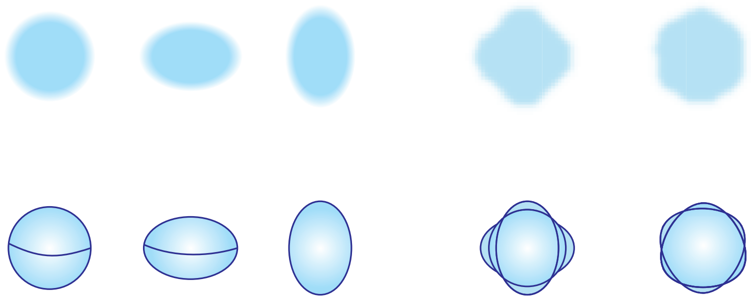


Figure 339 Various nuclear shapes – fixed (left) and oscillating (right), shown realistically as clouds (above) and simplified as geometric shapes (below)

process is similar to the emission of light by electrons in atoms. From the energy, the number and the lifetime of the excited states – they range from 1 ps to 300 d – researchers can deduce how the nucleons move inside the nucleus.

The photon energies define the 'colour' of the nucleus. It can be used, like all colours, to distinguish nuclei from each other and to study their motion. in particular, the colour of the γ-rays emitted by excited nuclei can be used to determine the chemical composition of a piece of matter. Some of these transition lines are so narrow that they can been used to study the change due to the chemical environment of the nucleus, to measure their motion or to detect the gravitational Doppler effect.

The study of γ-rays also allows to determine the shape of nuclei. Many nuclei are spherical; but many are prolate or oblate ellipsoids. Ellipsoids are favoured if the reduction in average electrostatic repulsion is larger than the increase in surface energy. All nuclei – except the lightest ones such as helium, lithium and beryllium – have a constant mass density at their centre, given by about 0.17 fermions per fm^3, and a skin thickness of about 2.4 fm, where their density decreases. Nuclei are thus small clouds, as shown in Figure 339.

We know that molecules can be of extremely involved shape. In contrast, nuclei are mostly spheres, ellipsoids or small variations of these. The reason is the short range, or better, the fast spatial decay of nuclear interactions. To get interesting shapes like in molecules, one needs, apart from nearest neighbour interactions, also next neighbour interactions and next next neighbour interactions. The strong nuclear interaction is too short ranged to make this possible. Or does it? It might be that future studies will discover that some nuclei are of more unusual shape, such as smoothed pyramids. Some predictions have been made in this direction; however, the experiments have not been performed yet. Ref. 897

The shape of nuclei does not have to be fixed; nuclei can also oscillate in shape. Such oscillations have been studied in great detail. The two simplest cases, the quadrupole and octupole oscillations, are shown in Figure 339. Obviously, nuclei can also rotate. Rapidly spinning nuclei, with a spin of up to $60\hbar$ and more, exist. They usually slow down step by step, emitting a photon and reducing their angular momentum at each step. Recently it was discovered that nuclei can also have bulges that rotate around a fixed core, a bit like Ref. 898

tides rotate around the Earth.

Motion in the nuclear domain – four types of motion

Nuclei are small because the nuclear interactions are short-ranged. Due to this short range, nuclear interactions play a role only in types of motion: scattering, bound motion, decay and a combination of these three called *nuclear reactions*. The history of nuclear physics showed that the whole range of observed phenomena can be reduced to these four fundamental processes. In each motion type, the main interest is what happens at the start and at the end; the intermediate situations are less interesting. Nuclear interactions thus lack the complex types of motion which characterize everyday life. That is the reason for the shortness of this chapter.

Scattering is performed in all accelerator experiments. Such experiments repeat for nuclei what we do when we look at an object. Seeing is a scattering process, as seeing is the detection of scattered light. Scattering of X-rays was used to see atoms for the first time; scattering of high energy alpha particles was used to discover and study the nucleus, and later the scattering of electrons with even higher energy was used to discover and study the components of the proton.

Bound motion is the motion of protons and neutrons inside nuclei or the motion of quarks inside hadrons. Bound motion determines shape and shape changes of compounds.

Decay is obviously the basis of radioactivity. Decay can be due to the electromagnetic, the strong or the weak nuclear interaction. Decay allows to study the conserved quantities of nuclear interactions.

Nuclear reactions are combinations of scattering, decay and possibly bound motion. Nuclear reactions are for nuclei what the touching of objects is in everyday life. Touching an object we can take it apart, break it, solder two objects together, throw it away, and much more. The same can be done with nuclei. In particular, nuclear reactions are responsible for the burning of the Sun and the other stars; they also tell the history of the nuclei inside our bodies.

Quantum theory showed that all four types of motion can be described in the same way. Each type of motion is due to the exchange of virtual particles. For example, scattering due to charge repulsion is due to exchange of virtual photons, the bound motion inside nuclei due to the strong nuclear interaction is due to exchange of virtual gluons, beta decay is due to the exchange of virtual W bosons, and neutrino reactions are due to the exchange of virtual Z bosons. The rest of this chapter explains these mechanisms in more details.

Nuclei react

Page 723

The first man who thought to have made transuranic elements, the Italian genius Enrico Fermi, received the Nobel prize for the discovery. Shortly afterwards, Otto Hahn and his collaborators Lise Meitner and Fritz Strassman showed that Fermi was wrong, and that his prize was based on a mistake. Fermi was allowed to keep his prize, the Nobel committee gave Hahn the Nobel prize as well, and to make the matter unclear to everybody and to women physicists in particular, the prize was not given to Lise Meitner. (After her death, a new element was named after her.)

When protons or neutrons were shot into nuclei, they usually remained stuck inside them, and usually lead to the transformation of an element into a heavier one. After having done this with all elements, Fermi used uranium; he found that bombarding it with neutrons, a new element appeared, and concluded that he had created a transuranic element. Alas, Hahn and his collaborators found that the element formed was well-known: it was barium, a nucleus with less than half the mass of uranium. Instead of remaining stuck as in the previous 91 elements, the neutrons had split the uranium nucleus. Hahn, Meitner and Strassmann had observed reactions such as:

$$^{235}\mathrm{U} + \mathrm{n} \rightarrow {}^{143}\mathrm{Ba} + {}^{90}\mathrm{Kr} + 3n + 170\,\mathrm{MeV} \,. \qquad (609)$$

Meitner called the splitting process *nuclear fission*. A large amount of energy is liberated in fission. In addition, several neutrons are emitted; they can thus start a chain reaction. Later, and (of course) against the will of the team, the discovery would be used to make nuclear bombs.

Reactions and decays are transformations. In each transformation, already the Greek taught us to search, first of all, for conserved quantities. Besides the well-known cases of energy, momentum, electric charge and angular momentum conservation, the results of nuclear physics lead to several new conserved quantities. The behaviour is quite constrained. Quantum field theory implies that particles and antiparticles (commonly denoted by a bar) must behave in compatible ways. Both experiment and quantum field theory show for example that every reaction of the type $\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D}$ implies that the reactions $\mathrm{A} + \overline{\mathrm{C}} \rightarrow \overline{\mathrm{B}} + \mathrm{D}$ or $\overline{\mathrm{C}} + \overline{\mathrm{D}} \rightarrow \overline{\mathrm{A}} + \overline{\mathrm{B}}$ or, if energy is sufficient, $\mathrm{A} \rightarrow \mathrm{C} + \mathrm{D} + \overline{\mathrm{B}}$, are also possible. Particles thus behave like conserved mathematical entities.

Experiments show that antineutrinos differ from neutrinos. In fact, all reactions confirm that the so-called *lepton number* is conserved in nature. The lepton number L is zero for nucleons or quarks, is 1 for the electron and the neutrino, and is −1 for the positron and the antineutrino.

In addition, all reactions conserve the so-called *baryon number*. The baryon number B is 1 for protons and neutrons (and 1/3 for quarks), and −1 for antiprotons and antineutrons (and thus −1/3 for antiquarks). So far, no process with baryon number violation has ever been observed. Baryon conservation is one reason for the danger of radioactivity, fission and fusion.

Bombs and nuclear reactors

Uranium fission is triggered by a neutron, liberates energy and produces several additional neutrons. It can trigger a chain reaction which can lead to an explosion or a controlled generation of heat. Once upon a time, in the middle of the twentieth century, these processes were studied by quite a number of researchers. Most of them were interested in making weapons or in using nuclear energy, despite the high toll these activities place on the economy, on human health and on the environment.

Most stories around this topic are absurd. The first nuclear weapons were built during the second world war with the smartest physicists that could be found. Everything was ready, including the most complex physical models, factories and an organization of incredible size. There was just one little problem: there was no uranium of sufficient quality. The mighty United States thus had to go around the world to shop for good uranium. They found it in the Belgian colony of Congo, in central Africa. In short, without the support of Belgium, which sold the Congolese uranium to the USA, there would have been no nuclear bomb, no early war end and no superpower status.

Congo paid a high price for this important status. It was ruled by a long chain of military dictators up to this day. But the highest price was paid by the countries that actually built nuclear weapons. Some went bankrupt, others remained underdeveloped, still other countries have amassed huge debts and have a large underprivileged population. There is no exception. The price of nuclear weapons has also been that some regions of our planet became uninhabitable, such as numerous islands, deserts and marine environments. But it could have been worse. When the most violent physicist ever, Edward Teller, made his first calculations about the hydrogen bomb, he predicted that the bomb would set the atmosphere into fire. Nobel prize winner Hans Bethe corrected the mistake and showed that nothing of this sort would happen. Nevertheless, the military preferred to explode the hydrogen bomb in the Bikini atoll, the most distant place from their homeland they could find. Today it is even dangerous simply to fly over that island. It was them noticed that nuclear test explosions increased ambient radioactivity in the atmosphere all over the world. Of the produced radioactive elements, ^{3}H is absorbed by humans in drinking water, ^{14}C and ^{90}Sr through food, and ^{137}Cs in both ways. In the meantime, all countries have agreed to perform their nuclear tests underground.

Ref. 901

But even peaceful nuclear reactors are dangerous. The reason was discovered in 1934 by Frédéric Joliot and his wife Irène, the daughter of Pierre and Marie Curie: *artificial radioactivity*. The Joliot–Curies discovered that materials irradiated by alpha rays become radioactive in turn. They found that alpha rays transformed aluminium into radioactive phosphorous:

$$^{27}_{13}\mathrm{Al} + {}^{4}_{2}\alpha \rightarrow {}^{4}_{15}\mathrm{P} \; . \tag{610}$$

In fact, almost all materials become radioactive when irradiated with alpha particles, neutrons or gamma rays. As a result, radioactivity itself can only be contained with difficulty. After a time which depends on the material and the radiation, the box that contains radioactive material has itself become radioactive.

The dangers of natural and artificial radioactivity are the reason for the high costs of nuclear reactors. After about thirty years of operation, reactors have to be dismantled. The radioactive pieces have to be stored in specially chosen, inaccessible places, and at the same time the workers' health must not be put in danger. The world over, many dismantlings are now imminent. The companies performing the job sell the service at high price. All operate in a region not far from the border to criminal activity, and since radioactivity cannot be detected by the human senses, many crossed it. In fact, an important nuclear reactor is (usually) not dangerous to humans: the Sun.

The Sun

Nuclear physics is the most violent part of physics. But despite this bad image, nuclear physics has something to offer which is deeply fascinating: the understanding of the Sun, the stars and the early universe.

Table 65 Some radioactivity measurements

Material	Activity in Bq/kg
air	*c.* 10^{-2}
sea water	10^{1}
human body	*c.* 10^{2}
cow milk	max. 10^{3}
pure ^{238}U metal	*c.* 10^{7}
highly radioactive α emitters	$> 10^{7}$
radiocarbon: ^{14}C (β emitter)	10^{8}
highly radioactive β and γ emitters	> 9
main nuclear fallout: ^{137}Cs, ^{90}Sr (α emitter)	$2 \cdot 10^{9}$
polonium, one of the most radioactive materials (α)	10^{24}

The Sun emits 385 YW of light. Where does this energy come from? If it came by burning coal, the Sun would stop burning after a few thousands of years. When radioactivity was discovered, researchers tested this possibility. However, even though radioactivity can produce more energy than chemical burning, the composition of the Sun – mostly hydrogen and helium – makes this impossible. In fact, the study of nuclei showed that the Sun burns by hydrogen *fusion*. Fusion is the composition of a large nucleus from smaller ones. In the Sun, the fusion reaction

$$4\,{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\,e^{+} + 2\,\nu + 4.4\,\mathrm{pJ} \qquad (611)$$

is the result of a continuous cycle of three separate nuclear reactions:

$$\begin{aligned} {}^{1}\mathrm{H} + {}^{1}\mathrm{H} &\rightarrow {}^{2}\mathrm{H} + e^{+} + \nu (\text{a weak nuclear reaction}) \\ {}^{2}\mathrm{H} + {}^{1}\mathrm{H} &\rightarrow {}^{3}\mathrm{He} + \gamma (\text{a strong nuclear reaction}) \\ {}^{3}\mathrm{He} + {}^{3}\mathrm{He} &\rightarrow {}^{4}\mathrm{He} + 2\,{}^{1}\mathrm{H} + \gamma\,. \end{aligned} \qquad (612)$$

In total, four protons are thus fused to one helium nucleus; if we include the electrons, four hydrogen atoms are fused to one helium atom with the emission of neutrinos and light with a total energy of 4.4 pJ (26.7 MeV). Most of the energy is emitted as light; around 10 % is carried away by neutrinos. The first of the three reaction of equation 612 is due to the weak nuclear interaction; this avoids that it happens too rapidly and ensures that the Sun will shine still for some time. Indeed, in the Sun, with a luminosity of 385 YW, there are thus about 10^{38} fusions per second. This allows to deduce that the Sun will last another handful of Ga (Gigayears) before it runs out of fuel. Ref. 899

The fusion reaction (612) takes place in the centre of the Sun. The energy carried away by the photons arrives at the Sun's surface about two hundred thousand years later; this delay is due to the repeated scattering of the photon by the constituents inside the Sun. After two-hundred thousand years, the photons take another 8.3 minutes to reach the

Table 66 Human exposure to radioactivity and the corresponding doses

Exposure	Dose
Daily human exposure:	
Average exposure to cosmic radiation in Europe, at sea level (3 km)	*c.* 0.3 mSv/a (1.2 mSv/a)
Average (and maximum) exposure to soil radiation, without radon	0.4 mSv/a (2 mSv/a)
Average (and maximum) inhalation of radon	1 mSv/a (100 mSv/a)
Average exposure due to internal radionuclides	0.3 mSv/a
natural content of ^{40}K in human muscles	10^{-4} Gy and 4500 Bq
natural content of Ra in human bones	$2 \cdot 10^{-5}$ Gy and 4000 Bq
natural content of ^{14}C in humans	10^{-5} Gy
Total average (and maximum) human exposure	2 mSv/a (100 mSv/a)
Common situations:	
Dental X-ray	*c.* 10 mSv equivalent dose
Lung X-ray	*c.* 0.5 mSv equivalent dose
Short one hour flight (see http://www.gsf.de/epcard)	*c.* 1 μSv
Transatlantic flight	*c.* 0.04 mSv
Maximum allowed dose at work	30 mSv/a
Deadly exposures:	
Ionization	0.05 C/kg can be deadly
Dose	100 Gy=100 J/kg is deadly in 1 to 3 days
Equivalent dose	more than 3 Sv/a leads to death

Earth and to sustain the life of all plants and animals.

Curiosities and fun challenges on radioactivity

▪ Researchers have developed an additional method to date stones using radioactivity. Whenever an alpha ray is emitted, the emitting atom gets a recoil. If the atom is part of a crystal, the crystal is damaged by the recoil. The damage can be seen under the microscope. By counting the damaged regions it is possible to date the time at which rocks have been crystallized. In this way it has been possible to determine when rocks formed from volcanic eruptions have been formed.

▪ It is still not clear whether the radiation of the Sun is constant over long time scales. There is an 11 year periodicity, the famous solar cycle, but the long term trend is still unknown. Precise measurements cover only the years from 1978 onwards, which makes only about 3 cycles. A possible variation of the solar constant might have important consequences for climate research; however, the issue is still open.

▪ Not all γ-rays are due to radioactivity. In the year 2000, an Italian group discovered

Ref. 895 that thunderstorms also emit γ-rays, of energies up to 10 MeV. The mechanisms are still being investigated.

▪ Chain reactions are quite common in nature. *Fire* is a chemical chain reaction, as are exploding fireworks. In both cases, material needs heat to burn; this heat is supplied by a neighbouring region that is already burning.

▪ Radioactivity can be extremely dangerous to humans. The best example is plutonium. Only 1 μg of this alpha emitter inside the human body are sufficient to cause lung cancer.

▪ Lead is slightly radioactive, because it contains the ^{210}Pb isotope, a beta emitter. For sensitive experiments, such as for neutrino experiments, radioactivity shields are need. The best material is lead, but obviously it has to be low radioactivity lead. Since the isotope ^{210}Pb has a half life of 22 years, one way to do it is to use old lead. In a precision neutrino experiment in the Gran Sasso, the research team uses lead from Roman times to reduce spurious signals.

▪ Not all reactors are human made. Natural reactors have been predicted in 1956 by Paul Kuroda. In 1972 the first example was found. In Oklo, in the African country of Gabon, there is a now famous geological formation where uranium is so common that two thousand million years ago a natural nuclear reactor has formed spontaneously – albeit a small one, with an estimated power generation of 100 kW. It has been burning for over 150 000 years, during the time when the uranium 235 percentage was 3% or more, as required for chain reaction. (Nowadays, the uranium 235 content on Earth is 0.7%.) The water of a nearby river was periodically heated to steam during an estimated 30 minutes; then the reactor cooled down again for an estimated 2.5 hours, since water is necessary to moderate the neutrons and sustain the chain reaction. The system has been studied in great detail, from its geological history up to the statements it makes about the constancy of the 'laws' of nature. The studies showed that 2000 million years ago the mechanisms were the same as those used today.

▪ High energy radiation is dangerous to humans. In the 1950s, when nuclear tests were still made above ground by the large armies in the world, the generals overruled the orders of the medical doctors. They positioned many soldiers nearby to watch the explosion, and worse, even ordered them to walk to the explosion site as soon as possible after the explosion. One does not need to comment on the orders of these generals. Several of these unlucky soldiers made a strange observation: during the flash of the explosion, they were able to see the bones in their own hand and arms. How can this be? Challenge 1342 ny

▪ The SI units for radioactivity are now common; in the old days, 1 Sv was called 100 rem or 'Röntgen equivalent man'; The SI unit for dose, $1\,\mathrm{Gy} = 1\,\mathrm{J/kg}$, replaces what used to be called 100 rd or Rad. The SI unit for exposition, 1 C/kg, replaces the older unit 'Röntgen', for which the relation is $1\,\mathrm{R} = 2.58 \cdot 10^{-4}\,\mathrm{C/kg}$.

▪ Nuclear bombs are terrible weapons. To experience their violence but also the criminal actions of many military people during the tests, have a look at the pictures of explosions. Ref. 896 Nuclear tests were made for many years in the times when generals refused to listen to doctors and scientists. Generals ordered to explode these weapons in the air, making the complete atmosphere of the world radioactive, hurting all mankind in doing so; worse, they even obliged soldiers to visit the radioactive explosion site, thus doing their best to let their own soldiers die from cancer and leukaemia.

30. The strong nuclear interaction and the birth of matter

Lernen ist Vorfreude auf sich selbst.*
Peter Sloterdijk

Since protons are positively charged, inside nuclei they must be bound by a force strong enough to keep them together against their electromagnetic repulsion. This is the strong nuclear interaction. Most of all, the strong interaction tells a good story about the stuff we are made of.

Why do the stars shine?

Don't the stars shine beautifully? I am the only person in the world who knows why they do.
Frits Houtermans (1903–1966)

All stars shine because of fusion. When two light nuclei are fused to a heavier one, some energy is set free, as the average nucleon is bound more strongly. This energy gain is possible until the nuclei of iron ^{56}Fe are made. For nuclei beyond this nucleus, the binding energies per nucleon then decrease again; thus fusion is not energetically possible.*

The different stars observed in the sky** can be distinguished by the type of fusion nuclear reaction that dominates. Most stars, in particular young or light stars run hydrogen fusion. In fact, there are at least two main types of hydrogen fusion: the direct hydrogen-hydrogen (p-p) cycle and the CNO cycle(s).

Page 854
The hydrogen cycle described above is the main energy source of the Sun. The simple description does not fully purvey the fascination of the process. On average, protons in the Sun's centre move with 600 km/s. Only if they hit each other precisely head-on can a nuclear reaction occur; in all other cases, the electrostatic repulsion between the protons keeps them apart. For an average proton, a head-on collision happens once every 7 thousand million years. Nevertheless, there are so many proton collisions in the Sun that every second four million tons of hydrogen are burned to helium.

Fortunately for us, the photons generated in the Sun's centre are 'slowed' down by the outer parts of the Sun. In this process, gamma photons are progressively converted to visible photons. As a result, the sunlight of today was in fact generated at the time of the Neandertalers, about 200 000 years ago. In other words, the effective speed of light right at the centre of the Sun is estimated to be around 10 km/year.

If a star has heavier elements inside it, the hydrogen fusion uses these elements as

* 'Learning is anticipated joy about yourself.'
* Thus fission becomes interesting as energy source for heavy nuclei.
** For the stars above you, see the http://me.in-berlin-de/~jd/himmel/himmel.00.11.html website.

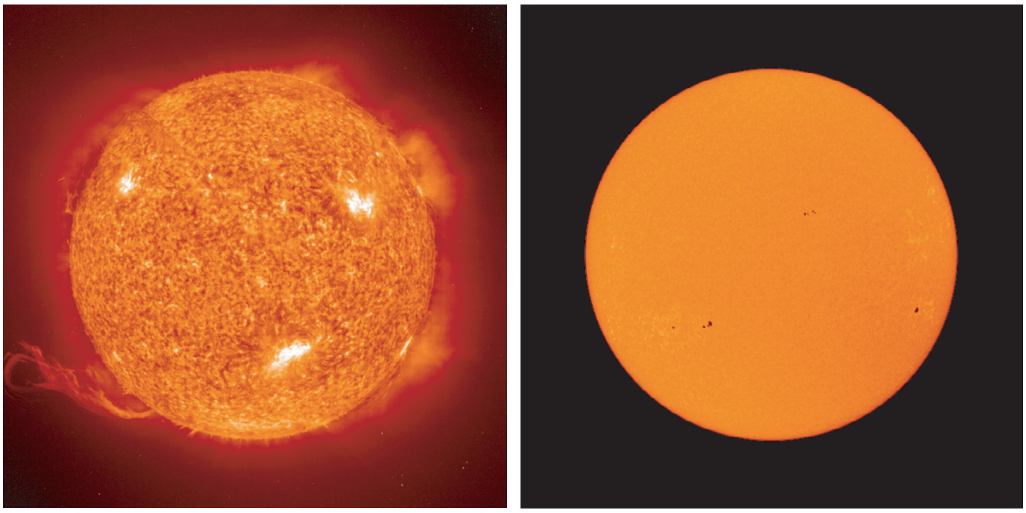


Figure 340 Photographs of the Sun at wavelengths of 30.4 nm (in the extreme ultraviolet, left) and around 677 nm (visible light, right, at a different date), by the soho mission (esa and nasa)

catalysts. This happens through the so-called CNO cycle, which runs as

$$\begin{aligned}
{}^{12}\mathrm{C} + {}^{1}\mathrm{H} &\rightarrow {}^{13}\mathrm{N} + \gamma \\
{}^{13}\mathrm{N} &\rightarrow {}^{13}\mathrm{C} + \mathrm{e}^{+} + \nu \\
{}^{13}\mathrm{C} + {}^{1}\mathrm{H} &\rightarrow {}^{14}\mathrm{N} + \gamma \\
{}^{14}\mathrm{N} + {}^{1}\mathrm{H} &\rightarrow {}^{15}\mathrm{O} + \gamma \\
{}^{15}\mathrm{O} &\rightarrow {}^{15}\mathrm{N} + \mathrm{e}^{+} + \nu \\
{}^{15}\mathrm{N} + {}^{1}\mathrm{H} &\rightarrow {}^{12}\mathrm{C} + {}^{4}\mathrm{He} \qquad (613)
\end{aligned}$$

The end result of the cycle is the same as that of the hydrogen cycle, both in nuclei and in energy. The CNO cycle is faster than hydrogen fusion, but requires higher temperatures, as the protons must overcome a higher energy barrier before reacting with carbon or nitrogen than when they react with another proton. (Why?) Due to the comparatively low temperature of a few tens of million kelvin inside the Sun, the CNO cycle is less important than the hydrogen cycle. (This is also the case for the other CNO cycles that exist.) These studies also explain why the Sun does not collapse. The Sun is a ball of hot gas, and the high temperature of its constituents prevents their concentration into a small volume. For some stars, the radiation pressure of the emitted photons prevents collapse; for others it is the Pauli pressure; for the Sun, like for the majority of stars, it is the usual thermal motion of the gas.

Challenge 1343 n

The nuclear reaction rates at the interior of a star are extremely sensitive to temperature. The carbon cycle reaction rate is proportional to between T^{13} for hot massive O stars and T^{20} for stars like the Sun. In red giants and supergiants, the triple alpha reaction rate is proportional to T^{40}; these strong dependencies imply that stars shine with constancy over medium times, since any change in temperature would be damped by a very efficient feedback mechanism. (Of course, there are exceptions: variable stars get brighter and darker with periods of a few days; and the Sun shows small oscillations in the minute range.)

How can the Sun's surface have a temperature of 6000 K, whereas the corona around it, the thin gas emanating from the Sun, reaches two million Kelvin? In the latter part of the twentieth century it was shown, using satellites, that the magnetic field of the Sun is the cause; through the violent flows in the Sun's matter, magnetic energy is transferred to the corona in those places were flux tubes form knots, above the bright spots in the left of Figure 340 or above the dark spots in the right photograph. As a result, the particles of the corona are accelerated and heat the whole corona.

When the Sun erupts, as shown in the lower left corner in Figure 340, matter is ejected far into space. When this matter reaches the Earth,* after being diluted by the journey, it affects the environment. Solar storms can deplete the higher atmosphere and can thus possibly trigger usual Earth storms. Other effects of the Sun are the formation of auroras and the loss of orientation of birds during their migration; this happens during exceptionally strong solar storms, as the magnetic field of the Earth is disturbed in these situations. The most famous effect of a solar storm was the loss of electricity in large parts of Canada in March of 1989. The flow of charged solar particles triggered large induced currents in the power lines, blew fuses and destroyed parts of the network, shutting down the power system. Millions of Canadians had no electricity, and in the most remote places it took two weeks to restore the electricity supply. Due to the coldness of the winter and a train accident resulting from the power loss, over 80 people died. In the meantime the network has been redesigned to withstand such events.

The proton cycle and the CNO cycles are not the only options. Heavier and older stars than the Sun can also shine through other fusion reactions. In particular, when hydrogen is consumed, such stars run *helium burning*:

$$3\,{}^{4}\mathrm{He} \rightarrow {}^{12}\mathrm{C}\ . \qquad (614)$$

This fusion reaction is of low probability, since it depends on three particles being at the same point in space at the same time. In addition, small amounts of carbon disappear rapidly via the reaction $\alpha + {}^{12}\mathrm{C} \rightarrow {}^{16}\mathrm{O}$. Nevertheless, since ${}^{8}\mathrm{Be}$ is unstable, the reaction with 3 alpha particles is the only way for the universe to produce carbon. All these negative odds are countered only by one feature carbon has an excited state at 7.65 MeV, which is 0.3 MeV above the sum of the alpha particle masses; the excited state resonantly enhances the low probability of the three particle reaction. Only in this way the universe is able to produce the atoms necessary for pigs, apes and people. The prediction of this resonance by Fred Hoyle is one of the few predictions in physics that used the simple experimental observation that humans exist. The story has lead to an huge outflow of metaphysical
Page 641 speculations, most of which are unworthy of being even mentioned.

* It might even be that the planets affect the solar wind; the issue is not settled and is still under study.

Where do our atoms come from?

People consist of electrons and various nuclei. Electrons, hydrogen and helium nuclei Page 427 are formed during the big bang. All other nuclei are formed in stars. Young stars run hydrogen burning or helium burning; heavier and older stars run neon-burning or even silicon-burning. These latter processes require high temperatures and pressures, which are found only in stars with a mass at least eight times that of the Sun. However, all fusion processes are limited by photodissociation and will not lead to nuclei heavier than ^{56}Fe.

Heavier nuclei can only be made by neutron capture. There are two main processes; the s-process (for 'slow') runs inside stars, and gradually builds up heavy elements until the most heavy, lead, from neutron flying around. The rapid r-process occurs in stellar explosions. Many stars die this violent death. Such an explosion has two main effects: on one hand it distributes most of the matter of the star, such carbon, nitrogen or oxygen, into space in the form of neutral atoms. On the other hand, new elements are synthesized during the explosion. The abundances of the elements in the solar system can be precisely measured. These several hundred data points correspond exactly with what is expected from the material ejected by a (type II) supernova explosion. In other words, the solar system formed from the remnants of a supernova, as did, somewhat later, life on Earth.* We all are recycled stardust.

The weak side of the strong interaction

Both radioactivity and medical images show that nuclei are composed. But quantum theory makes an additional prediction: protons and neutrons themselves must be composed. There are two reasons: nucleons have a finite size and their magnetic moments do not match the value predicted for point particles. The prediction of components inside the protons was confirmed in the late 1960s when Kendall, Friedman and Taylor shot high Ref. 888 energy electrons into hydrogen atoms. They found what that a proton contains three constituents with spin 1/2, which they called called *partons*. The experiment was able to 'see' the constituents through large angle scattering of electrons, in the same way that we see objects through large angle scattering of photons. These constituents correspond in number and properties to the so-called *quarks* predicted in the mid 1960s by Murray Gell-Ref. 889 Mann** and George Zweig.

It turns out that the interaction keeping the protons together in a nucleus, which was first described by Yukawa Hideki,*** is only a shadow of the interaction that keeps quarks

* By chance, the composition ratios between carbon, nitrogen and oxygen inside the Sun are the same as inside the human body.

** Murray Gell-Mann (b. 1929 New York, d.) Physics Nobel prize in 1969. He is the originator of the term 'quark'. (The term has two origins: officially, it is said to be taken from *Finnegan' Wake*, a novel by James Joyce; in reality, he took it from a German–Yiddish term meaning 'lean cheese' and used figuratively to mean 'silly idea'.)

Gell-Mann is a central figure of particle physics; he introduced the concept of strangeness, the renormalization group, the V-A interaction, the conserved vector current, the partially conserved axial current, the eightfold way, the quark model and quantum chromodynamics.

Gell-Mann is also known for his constant battle with Richard Feynman about who deserves to be called the most arrogant physicist of their university.

*** Yukawa Hideki (1907–1981), important Japanese physicist specialized in nuclear and particle physics.

figure to be inserted

Figure 341 A selection of mesons and baryons and their classification as bound states of quarks

together in a proton. Both are called by the same name. The two cases correspond somewhat to the two cases of electromagnetism found in atomic matter. Neon atoms show the cases most clearly: the strongest aspect of electromagnetism is responsible for the attraction of the electrons to the neon nuclei and its feebler 'shadow' is responsible for the attraction of neon atoms in liquid neon and for processes like evaporation. Both attractions are electromagnetic, but the strengths differ markedly. Similarly, the strongest aspect of the strong interaction leads to the formation of the proton and the neutron; the feeble aspect leads to the formation of nuclei and to alpha decay. Obviously, most can be learned by studying the strongest aspect.

Bound motion, the particle zoo and the quark model

Physicists are simple people. To understand the constituents of matter, and of nuclei in particular, they had no better idea than to take all particles they could get hold of and to smash them into each other. Many played this game for several decades.*

Imagine that you want to study how cars are built just by crashing them into each other. Before you get a list of all components, you must perform and study a non-negligible number of crashes. Most give the same result, and if you are looking for a particular part, you might have to wait for a long time. If the part is tightly attached to others, the crashes have to be especially energetic. Since quantum theory adds the possibility of transformations, reactions and excited states, the required diligence and patience is even greater than for car crashes. For many years, researchers collected an ever increasing number of debris. The list was overwhelming. Then came the quark model, which explained the whole mess as a consequence of only a few types of bound constituents.

Other physicists then added a few details and as a result, the whole list of debris could be ordered in tables such as the ones given in Figure 341 These tables were the beginning of the end of high energy physics. When the proton scattering experiments found that

He founded the journal Progress of Theoretical Physics and together with his class mate Tomonaga he was an example to many scientists in Japan. He received the 1949 Nobel prize for physics for this theory of mesons.

Challenge 1344 n
* In fact, quantum theory forbids any other method. Can you explain why?

protons are made of three constituents, the quark model became accepted all over the world.

The proton and the neutron are seen as combinations of two quarks, called *up* (u) and *down* (d). Later, other particles lead to the addition of four additional types of quarks. Their names are somewhat confusing: they are called *strange* (s), *charm* (c), *bottom* (b) – also called 'beauty' in the old days – and *top* (t) – called 'truth' in the past.

All quarks have spin one half; their electric charges are multiples of 1/3 of the electron charge. In addition, quarks carry a strong charge, which in modern terminology is called *colour*. In contrast to electromagnetism, which has only positive, negative, and neutral charges, the strong interaction has red, blue, green charges on one side, and anti-red, anti-blue and anti-green on the other. The neutral state is called 'white'. All baryons and mesons are white, in the same way that all atoms are neutral.

– CS – detail to be added – CS –

The mass, shape, and colour of protons

Frank Wilczek mentions that one of the main results of QCD, the theory of strong interactions, is to explain mass relations such as

Ref. 890

$$m_{\text{proton}} \sim e^{-k/\alpha} m_{\text{Planck}} \quad \text{and} \quad k = 11/2\pi \,,\; \alpha_{\text{unif}} = 1/25 \,. \tag{615}$$

Here, the value of the coupling constant α_{unif} is taken at the unifying energy, a factor of 1000 below the Planck energy. (See the section of unification below.) In other words, a general understanding of masses of bound states of the strong interaction, such as the proton, requires almost purely a knowledge of the unification energy and the coupling constant at that energy. The approximate value $\alpha_{\text{unif}} = 1/25$ is an extrapolation from the low energy value, using experimental data.

Page 873

The proportionality factor in expression (615) is still missing. Indeed, it is not easy to calculate. Many calculations are now done on computers. The most promising calculation simplify space-time to a lattice and then reduce QCD to lattice QCD. Using the most powerful computers available, these calculations have given predictions of the mass of the proton and other baryons within a few per cent.

Ref. 891

But the mass is not the only property of the proton. Being a cloud of quarks and gluons, it also has a shape. Surprisingly, it took a long time before people started to become interested in this aspect. The proton is made of two u quarks and one d quark. It thus resembles a ionized H_2^+ molecule, where one electron forms a cloud around two protons. Obviously, the H_2^+ molecule is elongated. Is that also the case for the proton? First results from 2003 seem to point into this direction.

The shape of a molecule will depend on whether other molecules surround it. Recent research showed that both the size and the shape of the proton in nuclei is slightly variable; both seem to depend on the nucleus in which the proton is built-in.

Ref. 892

Apart from shapes, molecules also have a colour. The colour of a molecule, like that of any object, is due to the energy absorbed when it is irradiated. For example, the H_2^+ molecule can absorb certain light frequencies by changing to an excited state. Protons

figure to be inserted

Figure 342 The spectrum of the excited states of proton and neutron

Ref. 893 and neutrons can also be excited; in fact, their excited states have been studied in detail; a summary is shown in Figure 342. It turns out that all these excitations can be explained as excited quarks states. For several excitations, the masses (or colours) have been calculated by lattice QCD to within 10%. The quark model and QCD thus structure and explain a large part of the baryon spectrum.

Obviously, in our everyday environment the energies necessary to excite nucleons do not appear – in fact, they do not even appear inside the Sun – and these excited states can be neglected. They only appear in particle accelerators. In a way, we can say that in our corner of the universe energies are to low to show the colour of protons.

Experimental consequences of the quark model

How can we pretend that quarks exist, even though they are never found alone? There are a number of arguments in favour.

- The quark model explains the non-vanishing magnetic moment of the neutron and explains the magnetic moments μ of the baryons. By describing the proton as a uud state and the neutron a udd state with no orbital angular momentum, we get

$$\mu_u = \tfrac{1}{5}(4\mu_p + \mu_n)\mu_0 \qquad \mu_d = \tfrac{1}{5}(4\mu_n + \mu_p)\mu_0 \qquad \text{where} \qquad \mu_0 = \hbar^2/Mc \qquad (616)$$

This means that $m_u = m_d = 330\,\text{MeV}$,a bit more than a third of the nucleon, whose mass is *c.* 940 MeV. If we assume that the quark magnetic moment is proportional to their charge, we predict a ratio of the magnetic moments of the proton and the neutron of $\mu_p/\mu_n = -1.5$; this prediction differs from measurements only by 3 %. Using the values for the magnetic moment of the quarks, magnetic moment values of over half a dozen of other baryons can be predicted. The results typically deviate from measurements by around 10 %; the sign is always correctly calculated.

- The quark model describes the quantum numbers of mesons and baryons. All texts on the quark model are full of diagrams such as those shown in Figure 341. These generalizations of the periodic table of the elements were filled in during the twentieth century; they allow a complete classification of all mesons and baryons as bound states of quarks.
- The quark model also explains the mass spectrum of baryons and mesons. The best

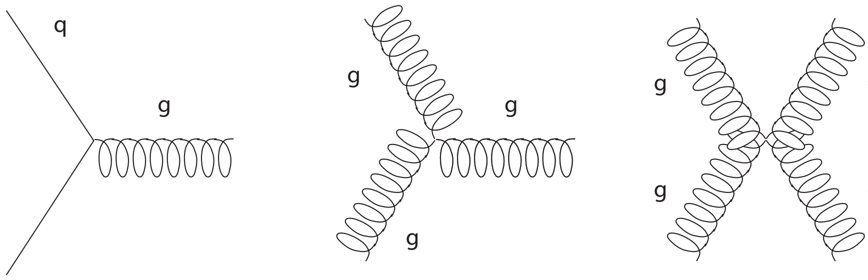


Figure 343 The central features of the qcd Lagrangian

predictions are made by lattice calculations. After one year of computer time, researchers were able to reproduce the masses of proton and neutron to within a few per cent. Even if one sets the u and d quark masses to zero, the resulting proton and neutron mass differ from experimental values only by 10 %. Ref. 903

The Lagrangian of quantum chromodynamics

All motion due to the strong interaction can be described by the three fundamental processes shown in Figure 343. A quark can emit or absorb a gluon, a gluon can emit another, and two gluons can scatter. In electrodynamics, only the first diagram is possible, in the strong interaction, the other two are added. Among others, the latter diagrams are responsible for the confinement of quarks.

All motion in the nuclear domain is described by its Lagrangian. We now know enough about the interaction to be able to understand the mathematical expression and to relate it to Figure 343. The Lagrangian density of quantum chromodynamics, often abbreviated QCD, is

$$\mathcal{L}_{QCD} = -\tfrac{1}{4}F^{(a)}_{\mu\nu}F^{(a)\mu\nu} - c^2\sum_q m_q \overline{\psi}^k_q \psi_{qk} + i\tfrac{\hbar}{c}c\sum_q \overline{\psi}^k_q \gamma^\mu (D_\mu)_{kl}\psi^l_q$$

where

$$\begin{aligned} F^{(a)}_{\mu\nu} &= \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_s f_{abc} A^b_\mu A^c_\nu \\ (D_\mu)_{kl} &= \delta_{kl}\partial_\mu - i g_s/2 \sum_a \lambda^a_{k,l} A^a_\mu \end{aligned} \qquad (617)$$

where $a = 1...8$ numbers the eight gluons, $k = 1, 2, 3$ numbers the three colours and $q = 1, ...6$ numbers the six quark flavours. The fields $\psi^k_q(x)$ are the fields of the quarks of flavour q and colour k, which are 4 component Dirac spinors, and $A^a_\mu(x)$ are the eight gluon fields. The quantities f_{abc} are the structure constants of the SU(3) algebra and the matrices $\lambda^a_{k,l}$ are a fundamental, 3-dimensional representation of the generators of the SU(3) algebra.*

* In their simplest form, the matrices γ_μ can be written as

$$\gamma_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \quad \text{and} \quad \gamma_n = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \quad \text{for } n = 1, 2, 3 \qquad (618)$$

The first term of the Lagrangian (617) represents the kinetic energy of the radiation (gluons), the second the kinetic energy of the matter particles (the quarks) and the third the interaction between the two. Gluons are massless; no gluon mass term appears in the Lagrangian.* Like in the whole of quantum field theory, also in this case the mathematical form of the Lagrangian is uniquely defined by the conditions of renormalizability, Lorentz invariance, gauge invariance (SU(3) in this case) and by specifying the different particles types (6 quarks in this case) with their masses and the coupling constant.

Only quarks and gluons appear in the Lagrangian of QCD, because only quarks and gluons interact via the strong force. This can be also expressed by saying that only quarks and gluons carry colour; *colour* is the source of the strong force in the same way that electric charge is the source of the electromagnetic field. In the same way as electric charge, colour charge is the source of the strong field; it is conserved in all interactions. Electric charge comes in two types, positive and negative; in a similar way, colour comes in three types, called red, green and blue. The neutral state, with no charge, is called white.

In opposition to electromagnetism, where the gauge group U(1) is abelian, the gauge

---

where the σ^i are the Pauli spin matrices.

* The matrices λ_a, $a = 1..8$, and the structure constants f_{abc} obey the relations

$$\begin{aligned}[\lambda_a, \lambda_b] &= 2i f_{abc} \lambda_c \\ \{\lambda_a, \lambda_b\} &= 4/3 \delta_{ab} I + 2 d_{abc} \lambda_c\end{aligned} \qquad (619)$$

where I is the unit matrix. The structure constants f_{abc}, which are odd under permutation of any pair of indices, and d_{abc}, which are even, are

abc	f_{abc}	abc	d_{abc}	abc	d_{abc}
123	1	118	$1/\sqrt{3}$	355	$1/2$
147	$1/2$	146	$1/2$	366	$-1/2$
156	$-1/2$	157	$1/2$	377	$-1/2$
246	$1/2$	228	$1/\sqrt{3}$	448	$-1/(2\sqrt{3})$
257	$1/2$	247	$-1/2$	558	$-1/(2\sqrt{3})$
345	$1/2$	256	$1/2$	668	$-1/(2\sqrt{3})$
367	$-1/2$	338	$1/\sqrt{3}$	778	$-1/(2\sqrt{3})$
458	$\sqrt{3}/2$	344	$1/2$	888	$-1/\sqrt{3}$
678	$\sqrt{3}/2$				

(620)

All other elements vanish. A fundamental 3-dimensional representation of the generators λ_a is given for example by the set of the *Gell-Mann matrices*

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$
$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} . \qquad (621)$$

There are eight matrices, one for each gluon type, with 3×3 elements, corresponding to the three colours of the strong interactions.

group SU(3) of the strong interactions is non-abelian. As a consequence, the colour field itself is charged, i.e. carries colour. Therefore gluons can interact with each other, in contrast to photons, which pass each other undisturbed.

We note that the Lagrangian contains parameters that remain unexplained:

(1) the number and the ('current') masses of the quarks are not explained by QCD.

(2) the coupling constant g_s of the strong interaction is unexplained. Often also the equivalent quantity $\alpha_s = g_s^2/4\pi$ is used to describe the coupling. Like for the case of the electroweak interactions, α_s and thus g_s depend on the energy Q of the experiment. This energy dependence is indeed observed in experiments; it is described by the renormalization procedure:

$$\alpha_s(Q^2) = \frac{12\pi}{33 - 2\mathrm{n}_f} \frac{1}{\ln \frac{Q^2}{\Lambda^2}} + \dots \qquad (622)$$

where n_f is the number of quarks with mass less than the energy scale Q and lies between 3 and 6. The strong coupling is thus completely described by the energy parameter $\Lambda \approx 0.25\,\mathrm{GeV}/c^2$. If α_s is known for one energy, it is known for all of them. Presently, the experimental value is $\alpha_s(Q^2 = 34\,\mathrm{GeV}) = 0.14 \pm 0.02$. Expression (622) also illustrates asymptotic freedom: α_s vanishes for high energies. In other words, at high energies quarks are freed from the strong interaction.*

At low energies, the coupling increases, and leads to quark confinement.** This behaviour is in contrast to the electroweak interactions, where the coupling increases with energy. We thus find that one parameter describing the strong coupling, Λ, remains unexplained and must be introduced into the Lagrangian from the beginning.

(3) The properties of space-time, its Lorentz invariance, its continuity and the number of its dimensions are obviously all unexplained and assumed from the outset.

The sizes and masses of quarks

The size of quarks, like that of all elementary particles, is predicted to be zero by quantum field theory. So far, no experiment has found an effect due to a finite quark size. Measurements show that quarks are surely smaller than 10^{-19} m. No definite size predictions have been made; quarks might have a size of the order of the grand unification scale, i.e. 10^{-32} m; however, so far this is speculation.

Ref. 904

Page 873

We noted in several places that a compound is always less massive than its components. But when the mass values for quarks are looked up in most tables, the masses of u and d quarks are only of the order of a few MeV/c^2, whereas the proton's mass is 938 MeV/c^2. What is the story here?

The definition of the masses for quarks is more involved than for other particles. Quarks are never found as free particles, but only in bound states. Quarks behave almost like free particles at high energies; this property is called *asymptotic freedom*. The mass of such a free quark is called *current quark mass*; for the light quarks it is only a few MeV/c^2.

* Asymptotic freedom was first discovered by Gerard 't Hooft; since he had the Nobel prize already, it the 2004 prize was then given to the next people who found it: David Gross, David Politzer and Frank Wilczek.

** Only at energies much larger than Λ can a perturbation expansion be applied.

At low energy, for example inside a proton, quarks are not free, but must carry along a large amount of energy due to the confinement process. As a result, bound quarks have a much larger *constituent quark mass*, which takes into account this confinement energy. To give an idea of the values, take a proton; the indeterminacy relation for a particle inside a sphere of radius 0.9 fm gives a momentum indeterminacy of around 190 MeV/c. In three dimensions this gives an energy of $\sqrt{3}$ times that value, or a mass of about 330 MeV/c^2. Three confined quarks are thus heavier than a proton, whose mass is 938 MeV/c^2; we can thus still say that a compound proton is less massive than its constituents. In short, the mass of the proton and the neutron is (almost exclusively) the kinetic energy of the quarks inside them, as their own rest mass is negligible. As Frank Wilczek says, some people put on weight even though they never eat anything heavy. Ref. 905

To complicate the picture, the distinction of the two mass types makes no sense for the top quark; this quark decays so rapidly that the confinement process has no time to set in. As a result, the top mass is again a mass of the type we are used to.

Confinement and the future of the strong interaction

The description of the proton mass using confined quarks should not hide the fact that the *complete* explanation of *quark confinement*, the lack of single quarks in nature, is the biggest challenge of theoretical high energy physics.

The Lagrangian of QCD differs from that of electromagnetism in a central aspect. So far, bound states cannot be deduced with a simple approximation method. In particular, the force dependence between two coloured particles, which does not decrease with increasing distance, but levels off at a constant value, does not follow directly from the Lagrangian. The constant value, which then leads to confinement, has been reproduced only in involved computer calculations.

In fact, the challenge is so tough that the brightest minds have been unable to solve it, so far. In a sense, it can be seen as the biggest challenge of all of physics, as its solution probably requires the unification of all interactions and most probably the unification with gravity. We have to leave this issue for later in our adventure.

Curiosities about the strong interactions

▪ The computer calculations necessary to extract particle data from the Lagrangian of quantum chromodynamics are among the most complex calculations ever performed. They beat weather forecasts, fluid simulations and the like by orders of magnitude. Nobody knows whether this is necessary: the race for a simple approximation method for finding solutions is still open.

▪ Even though gluons are massless, like photons and gravitons, there is no colour radiation. Gluons carry colour and couple to themselves; as a result, free gluons were predicted to directly decay into quark–antiquark pairs. This decay has indeed been observed in experiments at particle accelerators.

Something similar to colour radiation, but still stranger might have been found in 1997. First results seem to confirm the prediction of glueballs from numerical calculations. Ref. 906

▪ The latest fashion in high energy physics is the search for hybrid mesons, particles made of gluons and quarks. This fashion is not over yet; the coming years should settle Ref. 906

whether the candidates known so far really are hybrids.

▪ Do particles made of five quarks, so-called *pentaquarks*, exist? So far, they seem to exist only in a few laboratories in Japan, whereas other laboratories across the world fail to see them. The issue is still open.

▪ Whenever we look at a periodic table of the elements, we look at a manifestation of the strong interaction. The Lagrangian of the strong interaction describes the origin and properties of the presently known 115 elements.

Nevertheless a central aspect of nuclei is determined together with the electromagnetic interaction. Why are there around one hundred different elements? Because the electromagnetic coupling constant α is $1/137.036(1)$. Indeed, if the charge of a nucleus was much higher than around 130, the electric field around nuclei would lead to spontaneous electron–positron pair generation; the electron would fall into the nucleus and transform one proton into a neutron, thus inhibiting a larger proton number.

▪ To know more about radioactivity, its effects, its dangers and what a government can do about it, see the English and German language site of the Federal Office for Radiation Protection at http://www.bfs.de.

▪ From the years 1990 onwards, it has often been claimed that extremely poor countries are building nuclear weapons. Why is this not possible? Challenge 1345 n

▪ In the 1960s and 70s, it was discovered that the Sun pulsates with a frequency of 5 minutes. The effect is small, only 3 kilometres out of 1.4 million; still it is measurable. In the meantime, helioseismologists have discovered numerous additional oscillations of the Sun, and in 1993, even on other stars. Such oscillations allow to study what is happening inside stars, even separately in each of the layers they consist of.

▪ Historically, nuclear reactions also provided the first test of the relation $E = \gamma m c^2$. This was achieved in 1932 by Cockcroft and Walton. They showed that by shooting protons into lithium one gets the reaction

$$ {}^{7}_{3}\mathrm{Li} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{8}_{4}\mathrm{Be} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} + 17\,\mathrm{MeV}\ . \qquad (623) $$

The measured energy on the right is exactly the value that is derived from the differences in total mass of the nuclei on both sides.

▪ Some stars shine like a police siren: their luminosity increases and decreases regularly. Such stars, called *Cepheids*, are important because their period depends on their average (absolute) brightness. Measuring their period and their brightness on Earth thus allows astronomers to determine their distance.

31. The weak nuclear interaction and the handedness of nature

No interaction is as weird as the weak interaction. First of all, the corresponding 'weak radiation' consists of massive particles; there are two types, the neutral Z boson with a mass of 91.2 GeV – that is the mass of a silver atom – and the electrically charged W boson with a mass of 80.8 GeV. The masses are so large that free radiation exists only for an extremely short time, about 0.1 ys; then the particles decay. The large mass is the reason that the interaction is extremely short range and weak; any exchange of virtual particles scales with the negative exponential of the intermediate particle's mass.

The existence of a massive intermediate vector boson was already deduced in the 1940s; but theoretical physicists did not accept the idea until the Dutch physicist Gerard 't Hooft proved that it was possible to have such a mass without having problems in the rest of the theory. For this proof he later received the Nobel price of physics. Experimentally, the Z boson was found found as a virtual particle in 1973 and as a real particle in 1983, both times at CERN in Geneva. The last experiment was a year-long effort by thousands of people working together.

A central effect of the weak interaction is its ability to transform quarks. It is this property that is responsible for beta decay, where a d quark in a neutron is changed into a u quark, or for a crucial step in the Sun, where the opposite happens.

The next weird characteristic of the weak interaction is the nonconservation of parity under spatial inversion. The weak interaction distinguishes between mirror systems, in contrast to everyday life, gravitation, electromagnetism, and the strong interactions. Parity non-conservation had been predicted by 1956 by Lee and Yang, and was confirmed a few months later, earning them a Nobel prize. The most beautiful consequence of parity non-conservation property is its influence on the colour of certain atoms. This prediction was made in 1974 by Bouchiat and Bouchiat. The weak interaction is triggered by the weak charge of electrons and nuclei. Therefore, electrons in atoms do not exchange only virtual photons with the nucleus, but also virtual Z particles. The chance for this latter process is extremely small, around 10^{-11} times smaller than virtual photon exchange. But since the weak interaction is not parity conserving, this process allows electron transitions which are impossible by purely electromagnetic effects. In 1984, measurements confirmed that certain optical transitions of caesium atoms that are impossible via the electromagnetic interaction, are allowed when the weak interaction is taken into account. Several groups have improved these results and have been able to confirm the prediction of the weak interaction, including the charge of the nucleus, to within a few per cent.

Ref. 909

Ref. 910

Ref. 915

– CS – The section on weak interactions will be inserted here – CS –

Curiosities about the weak interactions

- The weak interaction is responsible for the burning of hydrogen to helium. Without helium, there would be no path to make still heavier elements. Thus we owe our own existence to the weak interaction.
- The weak interaction is required to have an excess of matter over antimatter. Without the parity breaking of the weak interactions, there would be no matter at all in the universe.

▪ Through the emitted neutrinos, the weak interaction helps to get the energy out of a supernova. If that were not the case, black holes would form, heavier elements – of which we are made – would not have been spread out into space, and we would not exist.

Ref. 908 ▪ The paper by Peter Higgs on the boson named after him is only 79 lines long, and has only five equations.

▪ The weak interaction is not parity invariant. In other words, when two electrons collide, the fraction of the collisions that happens through the weak interaction should behave differently than a mirror experiment. In 2004, polarized beams of electrons – either left-handed or right-handed – were shot at a matter target and the reflected electrons were counted. The difference was 175 parts per billion – small, but measurable. The experiment also confirmed the predicted weak charge of -0.046 of the electron.

▪ The weak interaction is also responsible for the heat produced inside the Earth. This heat keeps the magma liquid. As a result, the weak interaction, despite its weakness, is responsible for all earthquakes, tsunamis and volcanic eruptions.

▪ Beta decay, due to the weak interaction, separates electrons and protons. Only in 2005 people have managed to propose practical ways to use this effect to build long-life batteries that could be used in satellites. Future will tell whether the method will be successful.

Mass, the Higgs boson and a ten thousand million dollar lie

Difficile est satiram non scribere.

Juvenal*

In the years 1993 and 1994 an intense marketing campaign was carried out across the United States of America by numerous particle physicists. They sought funding for the 'superconducting supercollider', a particle accelerator with a circumference of 80 km. This should have been the largest human machine ever built, with a planned cost of more than ten thousand million dollars, aiming at finding the Higgs boson before the Europeans would do so, at a fraction of the cost. The central argument brought forward was the following: since the Higgs boson was the basis of mass, it was central to science to know about it. Apart from the discussion on the relevance of the argument, the worst is that it is wrong.

We have even seen that 95 % of the mass of protons, and thus of the universe, is due to confinement; it appears even if the quarks are approximated as massless. The Higgs boson is *not* responsible for the origin of mass itself; it just might shed some light on the issue. The whole campaign was a classic case of disinformation and many people involved have shown their lack of honesty. In the end, the project was stopped, mainly for financial reasons. But the disinformation campaign had deep consequences. US physicist lost their credibility. Even in Europe the budget cuts became so severe that the competing project in Geneva, though over ten times cheaper and financed by thirty countries instead of only one, was almost stopped as well. (Despite this hick-up, the project is now under way, scheduled for completion in 2007/8.)

* 'It is hard not to be satirical.' 1, 30

Neutrinium and other curiosities about the electroweak interaction

The weak interaction, with its breaking of parity and the elusive neutrino, exerts a deep
Ref. 907 fascination on all those who have explored it.

▪ Every second about 10^{16} neutrinos fly through our body. They are mainly created in the atmosphere by cosmic radiation, but also coming directly from the background radiation and from the centre of the Sun. Nevertheless, during our whole life – around 3 thousand million seconds – we have only a 10 % chance that one of them interacts with one of the $3 \cdot 10^{27}$ atoms of our body. The reason is that the weak interaction is felt only over distances less than 10^{-17} m, about 1/100th of the diameter of a proton. The weak interaction is indeed weak.

▪ The weak interaction is so weak that a neutrino–antineutrino annihilation – which is only possible by producing a massive intermediate Z boson – has never been observed up to this day.

▪ Only one type of particles interacts (almost) only weakly: neutrinos. Neutrinos carry no electric charge, no colour charge and almost no gravitational charge (mass). To get an impression of the weakness of the weak interaction, it is usually said that the probability of a neutrino to be absorbed by a lead screen of the thickness of one light-year is less than 50 %. The universe is thus essentially empty for neutrinos. Is there room for bound states of neutrinos circling masses? How large would such a bound state be? Can we imagine bound states, which would be called neutrinium, of neutrinos and antineutrinos circling each other? The answer depends on the mass of the neutrino. Bound states of massless particles do not exist. They could and would decay into two free massless particles.*

Since neutrinos are massive, a neutrino–antineutrino bound state is possible in prin-
Challenge 1346 ny ciple. How large would it be? Does it have excited states? Can they ever be detected? These issues are still open.

▪ Do ruminating *cows* move their jaws equally often in clockwise and anticlockwise direction? In 1927, the theoretical physicists Pascual Jordan and Ralph de Laer Kronig pub-
Ref. 911 lished a study showing that in Denmark the two directions are almost equally distributed. The rumination direction of cows is thus not related to the weak interaction.

▪ The weak interaction plays an important part in daily life. First of all, the Sun is shining. The fusion of two protons to deuterium, the first reaction of the hydrogen cycle,
Page 854 implies that one proton changes into a neutron. This transmutation and the normal beta
Challenge 1347 e decay have the same first-order Feynman diagram. The weak interaction is thus essential for the burning of the Sun. The weakness of the process is one of the guarantees that the Sun will continue burning for quite some time.

▪ Of course, the weak interaction is responsible for radioactive beta decay, and thus for part of the radiation background that leads to mutations and thus to biological evolution.

▪ What would happen if the Sun suddenly stopped shining? Obviously, temperatures would fall by several tens of degrees within a few hours. It would rain, and then all water would freeze. After four or five days, all animal life would stop. After a few weeks, the oceans would freeze; after a few months, air would liquefy.

▪ Not everything about the Sun is known. For example, the neutrino flux from the Sun oscillates with a period of 28.4 days. That is the same period with which the magnetic field of the Sun oscillates. The connections are still being studied.

* In particular, this is valid for photons bound by gravitation; this state is not possible.

▪ The energy carried away by neutrinos is important in supernovas; if neutrinos would not carry it away, supernovas would collapse instead of explode. That would have prevented the distribution of heavier elements into space, and thus our own existence.

▪ Even earlier on in the history of the universe, the weak interaction is important, as it prevents the symmetry between matter and antimatter, which is required to have an excess of one over the other in the universe.

▪ The beta decay of the radioactive carbon isotope ^{14}C has a decay time of 5568 a. This isotope is continually created in the atmosphere through the influence of cosmic rays (via the reaction $^{14}\mathrm{N} + \mathrm{n} \rightarrow \mathrm{p} + {}^{14}\mathrm{C}$), so that its concentration in air is thus (mostly) constant over time. Inside living plants, the metabolism thus (unwillingly) maintains the same concentration. In dead plants, the decay sets in. The decay time of a few thousand years is particularly useful to date historic material. The method, called radiocarbon dating, was used to determine the age of mummies, the age of prehistoric tools and the age of religious relics. Several relics turned out to be forgeries, such as a cloth in Turin, and several of their wardens turned out to be crooks.

▪ Due to the large toll it placed on society, research in nuclear physics, like poliomyelitis, has almost disappeared from the planet. Like poliomyelitis, nuclear research is kept alive only in a few highly guarded laboratories around the world, mostly by questionable figures, in order to build dangerous weapons. Only a small number of experiments carried on by a few researchers are able to avoid this involvement and continue to advance the topic.

▪ Interesting aspects of nuclear physics appear when powerful lasers are used. In 1999, a British team led by Ken Ledingham observed laser induced uranium fission in ^{238}U nuclei. In the meantime, this has even be achieved with table-top lasers. The latest feat, in 2003, was the transmutation of ^{129}I to ^{128}I with a laser. This was achieved by focussing
Ref. 912
a 360 J laser pulse onto a gold foil; the ensuing plasma accelerates electrons to relativistic speed, which hit the gold and produce high energy γ rays that can be used for the transmutation.

32. The standard model of elementary particle physics – as seen on television

– CS – the section will appear in the next version – CS –

Conclusion and open questions about the standard model

The standard model clearly distinguishes elementary from composed particles. It provides the full list of properties that characterizes a particle and thus any object in nature: charge, spin, isospin, parity, charge parity, strangeness, charm, topness, beauty, lepton number, baryon number and mass. The standard model also describes interactions as exchange of virtual radiation particles. It describes the types of radiation that are found in nature at experimentally accessible energy. In short, the standard model realizes the dream of the ancient Greeks, plus a bit more: we have the bricks that compose all of matter and radiation, and in addition we know precisely how they move and interact.

But we also know what we still do not know:

- we do not know the origin of the coupling constants;
- we do not know the origin of the symmetry groups;
- we do not know the details of confinement;
- we do not know whether the particle concept survives at high energy;
- we do not know what happens in curved space-time.

To study these issues, the simplest way is to explore nature at particle energies that are as high as possible. There are two methods: building large experiments or making some calculations. Both are important.

33. Grand unification – a simple dream

> Materie ist geronnenes Licht.*
>
> Albertus Magnus

Is there a common origin of the three particle interactions? We have seen in the preceding sections that the Lagrangians of the electromagnetic, the weak and the strong nuclear interactions are determined almost uniquely by two types of requirements: to possess a certain symmetry and to possess mathematical consistency. The search for *unification* of the interactions thus requires the identification of th unified symmetry of nature. In recent decades, several candidate symmetries have fuelled the hope to achieve this program: grand unification, supersymmetry, conformal invariance and coupling constant duality. The first of them is conceptually the simplest.

At energies below 1000 GeV there are no contradictions between the Lagrangian of the standard model and observation. The Lagrangian looks like a low energy approximation. It should thus be possible (attention, this a belief) to find a unifying symmetry that *contains* the symmetries of the electroweak and strong interactions as subgroups and thus as different aspects of a single, unified interaction; we can then examine the physical properties that follow and compare them with observation. This approach, called *grand unification*, attempts the unified description of all types of matter. All known elementary particles are seen as fields which appear in a Lagrangian determined by a single symmetry group.

Like for each gauge theory described so far, also the grand unified Lagrangian is mainly determined by the symmetry group, the representation assignments for each particle, and

* 'Matter is coagulated light.' Albertus Magnus (b. *c.* 1192 Lauingen, d. 1280 Cologne), the most important thinker of his time.

the corresponding coupling constant. A general search for the symmetry group starts with all those (semisimple) Lie groups which contain $U(1)\times SU(2)\times SU(3)$. The smallest groups with these properties are SU(5), SO(10) and E(8); they are defined in Appendix D. For each of these candidate groups, the experimental consequences of the model must be studied and compared with experiment.

Ref. 913

Experimental consequences

Grand unification makes several clear experimental predictions.

- Any grand unified model predicts relations between the quantum numbers of all elementary particles – quarks and leptons. As a result, grand unification explains why the electron charge is exactly the opposite of the proton charge.
- Grand unification predicts a value for the weak mixing angle θ_{W} that is not determined by the standard model. The predicted value,

$$\sin^2\theta_{W,th} = 0.2 \tag{624}$$

is close to the measured value of

$$\sin^2\theta_{W,ex} = 0.231(1) . \tag{625}$$

- All grand unified models predict the existence of magnetic monopoles, as was shown by Gerard 't Hooft. However, despite extensive searches, no such particles have been found yet. Monopoles are important even if there is only one of them in the whole universe: the existence of a single monopole implies that electric charge is quantized. Grand unification thus explains why electric charge appears in multiples of a smallest unit.

Ref. 916

- Grand unification predicts the existence of heavy intermediate vector bosons, called *X bosons*. Interactions involving these bosons do not conserve baryon or lepton number, but only the difference $B-L$ between baryon and lepton number. To be consistent with experiment, the X bosons must have a mass of the order of 10^{16} GeV.
- Most spectacularly, the X bosons grand unification implies that the proton decays. This prediction was first made by Pati and Salam in 1974. If protons decay, means that neither coal nor diamond* – nor any other material – is for ever. Depending on the precise symmetry group, grand unification predicts that protons decay into pions, electrons, kaons or other particles. Obviously, we know 'in our bones' that the proton lifetime is rather high, otherwise we would die of leukaemia; in other words, the low level of cancer already implies that the lifetime of the proton is larger than 10^{16} a.

Detailed calculations for the proton lifetime τ_p using SU(5) yield the expression

$$\tau_{p} \approx \frac{1}{\alpha_G^2(M_X)}\frac{M_X^4}{M_p^5} \approx 10^{31\pm1}\,\mathrm{a} \tag{626}$$

where the uncertainty is due to the uncertainty of the mass M_X of the gauge bosons in-

* As is well known, diamond is not stable, but metastable; thus diamonds are not for ever, but coal might be, if protons do not decay.

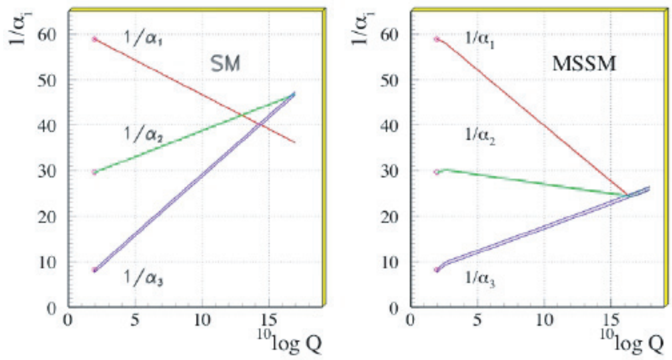


Figure 344 The behaviour of the three coupling constants with energy for the standard model (left) and for the minimal supersymmetric model (right) (© Dmitri Kazakov)

volved and to the exact decay mechanism. Several large experiments aim to measure this lifetime. So far, the result is simple but clear. Not a single proton decay has ever been observed. The experiments can be summed up by

Ref. 917

$$\begin{aligned} \tau(\mathrm{p} \to e^+\,\pi^0) &> 5\cdot 10^{33}\,\mathrm{a} \\ \tau(\mathrm{p} \to K^+\,\bar{\nu}) &> 1.6\cdot 10^{33}\,\mathrm{a} \\ \tau(\mathrm{n} \to e^+\,\pi^-) &> 5\cdot 10^{33}\,\mathrm{a} \\ \tau(\mathrm{n} \to K^0\,\bar{\nu}) &> 1.7\cdot 10^{32}\,\mathrm{a} \end{aligned} \qquad (627)$$

These values are higher than the prediction by SU(5). To settle the issue definitively, one last prediction of grand unification remains to be checked: the unification of the coupling constants.

The state of grand unification

The estimates of the grand unification energy are near the Planck energy, the energy at which gravitation starts to play a role even between elementary particles. As grand unification does not take gravity into account, for a long time there was a doubt whether something was lacking in the approach. This doubt changed into certainty when the precision measurements of the coupling constants became available. This happened in 1991, when these measurements were shown as Figure 344. It turned out that the SU(5) prediction of the way the constants evolve with energy imply that the three constants do *not* meet at the grand unification energy. Simple grand unification by SU(5) is thus definitively ruled out.

Ref. 918

This state of affairs is changed if supersymmetry is taken into account. Supersymmetry is the low-energy effect of gravitation in the particle world. Supersymmetry predicts new particles that change the curves at intermediate energies, so that they all meet at a grand unification energy of about 10^{16} GeV. The inclusion of supersymmetry also puts the proton lifetime prediction back to a value higher (but not by much) than the present experimental bound and predicts the correct value of the mixing angle. With supersymmetry, we can thus retain all advantages of grand unification (charge quantization, fewer parameters) without being in contradiction with experiments. The predicted particles, not yet found, are in a region accessible to the LHC collider presently being built at CERN in Geneva. We will explore supersymmetry later on.

Eventually, some decay and particle data will become available. Even though these experimental results will require time and effort, a little bit of thinking shows that they probably will be only partially useful. Grand unification started out with the idea to unify the description of matter. But this ambitious goal *cannot* been achieved in this way. Grand unification does eliminate a certain number of parameters from the Lagrangians of QCD and QFD; on the other hand, some parameters remain, even if supersymmetry is added. Most of all, the symmetry group must be put in from the beginning, as grand unification cannot deduce it from a general principle.

If we look at the open points of the standard model, grand unification reduces their number. However, grand unification only *shifts* the open questions of high energy physics to the next level, while keeping them unanswered. Grand unification remains a low energy *effective* theory. Grand unification does not tell us what elementary particles are; the name 'grand unification' is ridiculous. In fact, the story of grand unification is a first hint that looking at higher energies using only low-energy concepts is not the way to solve the mystery of motion. We definitively need to continue our adventure.

Bibliography

885 A excellent technical introduction to nuclear physics is Bogdan Povh, Klaus Rith, Christoph Scholz & Frank Zetsche, *Teilchen und Kerne*, Springer, 5th edition, 1999. It is also available in English translation.

One of the best introductions into particle physics is Kurt Gottfried & Victor F. Weisskopf, *Concepts of Particle Physics*, Clarendon Press, Oxford, 1984. Victor Weisskopf was one of the heroes of the field, both in theoretical research and in the management of CERN, the particle research institution. Cited on page 836.

886 W.C.M. Weijmar Schultz & al., Magnetic resonance imaging of male and female genitals during coitus and female sexual arousal, *British Medical Journal* **319**, pp. 1596–1600, December 18, 1999, available online as http://www.bmj.com/cgi/content/full/319/7225/1596. Cited on page 837.

887 A good overview is given by A.N. Halliday, Radioactivity, the discovery of time and the earliest history of the Earth, *Contemporary Physics* **38**, pp. 103–114, 1997. Cited on page 848.

888 Kendall, Friedman and Taylor received the 1990 Nobel prize for Physics for a series of experiments they conducted in the years 1967 to 1973. The story is told in the three Nobel lectures R.E. Taylor, Deep inelastic scattering: the early years, *Review of Modern Physics* **63**, pp. 573–596, 1991, H.W. Kendall, Deep inelastic scattering: Experiments on the proton and the observation of scaling, *Review of Modern Physics* **63**, pp. 597–614, 1991, and J.I. Friedman, Deep inelastic scattering: Comparisons with the quark model, *Review of Modern Physics* **63**, pp. 615–620, 1991. Cited on page 860.

889 G. Zweig, An SU3 model for strong interaction symmetry and its breaking II, CERN Report No. 8419TH. 412, February 21, 1964. Cited on page 860.

890 F. Wilczek, Getting its from bits, *Nature* **397**, pp. 303–306, 1999. Cited on page 862.

891 For an overview of lattice QCD calculations, see ... Cited on page 862.

892 S. Strauch & al., Polarization transfer in the ^{4}He (e,e'p) ^{3}H reaction up to Q^2 = 2.6(GeV,c)^2/ *Physical Review Letters* **91**, p. 052301, 2003. Cited on page 862.

893 The excited states of the proton and the neutron can be found in on the particle data group website at http://pdg.web.cern.ch. Cited on page 863.

894 An older but fascinating summary of solar physics is R. Kippenhahn, *Hundert Milliarden Sonnen*, Piper, 1980. which is available also in English translation. No citations.

895 M. Brunetti, S. Cecchini, M. Galli, G. Giovannini & A. Pagliarin, Gamma-ray bursts of atmospheric origin in the MeV energy range, *Geophysical Research Letters* **27**, p. 1599, 1 June 2000. Cited on page 856.

896 A book with nuclear explosion photographs is Michael Light, *100 Suns*, Jonathan Cape, 2003. Cited on page 856.

897 J. Dudek, A. God, N. Schunck & M. Mikiewicz, Nuclear tetrahedral symmetry: possibly present throughout the periodic table, *Physical Review Letters* **88**, p. 252502, 24 June 2002. Cited on page 850.

898 A good introduction is R. Clark & B. Wodsworth, A new spin on nuclei, *Physics World* pp. 25–28, July 1998. Cited on page 850.

899 M. Nauenberg & V.F. Weisskopf, Why does the sun shine?, *American Journal of Physics* **46**, pp. 23–31, 1978. Cited on page 854.

900 R.A. Alpher, H. Bethe & G. Gamow, The Origin of Chemical Elements, *Physical Review* 73, p. 803-804, 1948. No citations.

901 John Horgan, *The End of Science – Facing the Limits of Knowledge in the Twilight of the Scientific Age*, Broadway Books, 1997, chapter 3, note 1. Cited on page 853.

902 M. Chantell, T.C. Weekes, X. Sarazin & M. Urban, Antimatter and the moon, *Nature* **367**, p. 25, 1994. M. Amenomori & al., Cosmic ray shadow by the moon observed with the Tibet air shower array, Proceedings of the 23rd International Cosmic Ray Conference, Calgary **4**, pp. 351–354, 1993. M. Urban & al., *Nuclear Physics, Proceedings Supplement* **14B**, pp. 223–236, 1990. Cited on page 845.

903 See also C. Bernard & al., Light hadron spectrum with Kogut–Susskind quarks, *Nuclear Physics, Proceedings Supplement* **73**, p. 198, 1999, and references therein. Cited on page 864.

904 F. Abe & al., Measurement of dijet angular distributions by the collider detector at Fermilab, *Physical Review Letters* **77**, pp. 5336–5341, 1996. Cited on page 866.

905 The approximation of QCD with zero mass quarks is described by F. Wilczek, Getting its from bits, *Nature* **397**, pp. 303–306, 1999. Cited on page 867.

906 F. Close, Glueballs and hybrids: new states of matter, *Contemporary Physics* **38**, pp. 1–12, 1997. Cited on page 867.

907 K. Grotz & H.V. Klapdor, *Die schwache Wechselwirkung in Kern-, Teilchen- und Astrophysik*, Teubner Verlag, Stuttgart, 1989. Also available in English and in several other languages. Cited on page 871.

908 P.W. Higgs, Broken symmetries, massless particles and gauge fields, *Physics Letters* **12**, pp. 132–133, 1964. He then expanded the story in P.W. Higgs, Spontaneous symmetry breakdown without massless bosons, *Physical Review* **145**, p. 1156-1163, 1966. Higgs gives most credit to Anderson, instead of to himself; he also mentions Brout and Englert, Guralnik, Hagen, Kibble and 't Hooft. Cited on page 870.

909 M.A. Bouchiat & C.C. Bouchiat, Weak neutral currents in atomic physics, *Physics Letters B* **48**, pp. 111–114, 1974. U. Amaldi, A. Böhm, L.S. Durkin, P. Langacker, A.K. Mann, W.J. Marciano, A. Sirlin & H.H. Williams, Comprehensive analysis of data pertaining to the weak neutral current and the intermediate-vector-boson masses, *Physical Review D* **36**, pp. 1385–1407, 1987. Cited on page 869.

910 M.C. Noecker, B.P. Masterson & C.E. Wiemann, Precision measurement of parity nonconservation in atomic cesium: a low-energy test of electroweak theory, *Physical Review Letters* **61**, pp. 310–313, 1988. See also D.M. Meekhof & al., High-precision measurement of parity nonconserving optical rotation in atomic lead, *Physical Review Letters* **71**, pp. 3442–3445, 1993. Cited on page 869.

911 Rumination is studied in P. Jordan & R. de Laer Kronig, in *Nature* **120**, p. 807, 1927. Cited on page 871.

912 K.W.D. Ledingham & al., Photonuclear physics when a multiterawatt laser pulse interacts with solid targets, *Physical Review Letters* **84**, pp. 899–902, 2000. K.W.D. Ledingham & al., Laser-driven photo-transmutation of Iodine-129 – a long lived nuclear waste product, *Journal of Physics D: Applied Physics* **36**, pp. L79–L82, 2003. R.P. Singhal, K.W.D. Ledingham & P. McKenna, Nuclear physics with ultra-intense lasers – present status and future prospects, *Recent Research Developments in Nuclear Physics* **1**, pp. 147–169, 2004. Cited on page 872.

913 For the bibliographic details of the last Review of particle physics, see Appendix B. The present status on grand unification can also be found in the respective section of the overview. Cited on page 874.

914 J. Tran Thanh Van, ed., *CP violation in Particle Physics and Astrophysics*, Proc. Conf. Chateau de Bois, France, May 1989, editions Frontières, Gif sur Ivette, 1990. No citations.

915 S.C. Bennet & C.E. Wiemann, Measurement of the 6S – 7S transition polarizability in atomic cesium and an improved test of the standard model, *Physical Review Letters* **82**, pp. 2484–2487, 1999. The group has also measured the spatial distribution of the weak charge, the so-called the *anapole moment*; see C.S. Wood & al., Measurement of parity nonconservation and an anapole moment in cesium, *Science* **275**, pp. 1759–1763, 1997. Cited on page 869.

916 H. Jeon & M. Longo, Search for magnetic monopoles trapped in matter, *Physical Review Letters* **75**, pp. 1443–1447, 1995. Cited on page 874.

917 On proton decay rates, see the data of the particle data group, at http://pdg.web.cern.ch. Cited on page 875.

918 U. Amaldi, W. de Boer & H. Fürstenau, Comparison of grand unified theories with elektroweak and strong coupling constants measured at LEP, *Physics Letters* **260**, pp. 447–455, 1991. This widely cited paper is the standard reference for this issue. Cited on page 875.

Chapter IX

Advanced Quantum Theory (Not yet Available)

– CS – this chapter will be made available in the future – CS –

Chapter X

Quantum Physics in a Nutshell

Quantum theory's essence: the lack of the infinitely small

Compared to classical physics, quantum theory is remarkably more complex. The basic idea however, is simple: in nature there is a minimum change, or a minimum action, or again, a minimum angular momentum $\hbar/2$. The minimum action leads to all the strange observations made in the microscopic domain, such as wave behaviour of matter, tunnelling, indeterminacy relations, randomness in measurements, quantization of angular momentum, pair creation, decay, indistinguishability and particle reactions. The mathematics is often disturbingly involved. Was this part of the walk worth the effort? It was. The accuracy is excellent and the results profound. We give an overview of both and then turn to the list of questions that are still left open.

Achievements in precision

Quantum theory improved the accuracy of predictions from the few – if any – digits common in classical mechanics to the full number of digits – sometimes fourteen – that can be measured today. The limited precision is usually *not* given by the inaccuracy of theory, it is given by the measurement accuracy. In other words, the agreement is only limited by the amount of money the experimenter is willing to spend. Table 67 shows this in more detail.

Table 67 Some comparisons between classical physics, quantum theory and experiment

Observable	Classical prediction	Prediction of quantum theory[a]	Measurement	Cost estimate[b]
Simple motion of bodies				
Indeterminacy	0	$\Delta x \Delta p \geqslant \hbar/2$	$(1 \pm 10^{-2})\,\hbar/2$	10 k€
Wavelength of matter beams	none	$\lambda p = 2\pi\hbar$	$(1 \pm 10^{-2})\,\hbar$	10 k€
Tunnelling rate in alpha decay	0	$\tau = ...$	$(1 \pm 10^{-2})\,\tau$	0.5 M€
Compton wavelength	none	$\lambda_c = h/m_e c$	$(1 \pm 10^{-3})\,\lambda$	20 k€
Pair creation rate	0	...	...	20 M€
Radiative decay time in hydrogen	none	$\tau \sim 1/n^3$	...	5 k€

Observable	Classical prediction	Prediction of quantum theory[a]	Measurement	Cost[b] estimate
Smallest action and angular momentum	0	$\hbar/2$	$(1 \pm \pm 10^{-6})\,\hbar/2$	10 k€
Casimir effect	0	$ma/A = (\pi^2\hbar c)/(240 r^4)$	$(1 \pm 10^{-3})\,ma$	30 k€
Colours of objects				
Lamb shift	none	$\Delta\lambda = 1057.86(1)$ MHz	$(1 \pm 10^{-6})\,\Delta\lambda$	50 k€
Rydberg constant	none	$R_\infty = m_e c\alpha^2/2h$	$(1 \pm 10^{-9})\,R_\infty$	50 k€
Stefan–Boltzmann constant	none	$\sigma = \pi^2 k^4/60\hbar^3 c^2$	$(1 \pm 3 \cdot 10^{-8})\,\sigma$	20 k€
Wien displacement constant	none	$b = \lambda_{max} T$	$(1 \pm 10^{-5})\,b$	20 k€
Refractive index of ...	none	...	...	...
Photon-photon scattering	0	...	...	50 M€
Particle and interaction properties				
Electron gyromagnetic ratio	1 or 2	2.002 319 304 3(1)	2.002 319 304 3737(82)	30 M€
Z boson mass	none	$m_Z^2 = m_W^2(1 + \sin\theta_W^2)$	$(1 \pm 10^{-3})\,m_Z$	100 M€
proton mass	none	$(1 \pm 5\,\%)\,m_p$	m_p =1.67 yg	1 M€
reaction rate	0	...	...	
Composite matter properties				
Atom lifetime	≈ 1 µs	∞	$> 10^{20}$ a	10 k€
Molecular size	none	from QED	within 10^{-3}	20 k€
Von Klitzing constant	∞	$h/e^2 = \mu_0 c/2\alpha$	$(1 \pm 10^{-7})\,h/e^2$	1 M€
AC Josephson constant	0	$2e/h$	$(1 \pm 10^{-6})\,2e/h$	5 M€
Heat capacity of metals at 0 K	0	25 J/K	$< 10^{-3}$ J/K	10 k€
Water density	none	...	1000 kg/m^3	10 k€
Minimum electr. conductivity	0	$G = 2e^2/\hbar$	$G(1 \pm 10^{-3})$	3 k€
Proton lifetime	≈ 1 µs	∞	$> 10^{35}$ a	100 M€

Page 887
Page 1060
a. All these predictions are calculated from the quantities of Table 68, and no other input. Their most precise experimental values are given in Appendix B.

Challenge 1348 n
b. Sometimes the cost for the calculation of the prediction is higher than that of its measurement. (Can you spot the examples?) The sum of the two is given.

We notice that the predicted values are not noticeably different from the measured ones. If we remember that classical physics does not allow to calculate any of the predicted values we get an idea of the progress quantum physics has allowed. But despite this im-

Page 887 pressive agreement, there still are *unexplained* observations. In fact, these unexplained observations provide the input for the calculations just cited; we list them in detail below, in Table 68.

In summary, in the microscopic domain we are left with the impression that quantum theory is in *perfect* correspondence with nature; despite prospects of fame and riches, despite the largest number of researchers ever, no contradiction with observation has been found yet.

Physical results of quantum theory

Deorum offensae diis curae.
Voltaire, *Traité sur la tolérance.*

All of quantum theory can be resumed in two sentences.

▷ *In nature, actions smaller than* $\hbar/2 = 0.53 \cdot 10^{-34}$ Js *are not observed.*

▷ *All intrinsic properties in nature – with the exception of mass – such as electric charge, spin, parities, etc., appear as integer numbers; in composed systems they either add or multiply.*

The second statement in fact results from the first. The existence of a smallest action in nature directly leads to the main lesson we learned about motion in the second part of our adventure:

▷ *If it moves, it is made of particles.*

This statement applies to everything, thus to all objects and to all images, i.e. to matter and to radiation. Moving stuff is made of *quanta*. Stones, water waves, light, sound waves, earthquakes, gelatine and everything else we can interact with is made of particles. We started the second part of our mountain ascent with the title question: what is matter and what are interactions? Now we know: they are composites of elementary particles.

To be clear, an *elementary particle* is a countable entity, smaller than its own Compton wavelength, described by energy, momentum, and the following *complete* list of intrinsic properties: mass, spin, electric charge, parity, charge parity, colour, isospin, strangeness, charm, topness, beauty, lepton number, baryon number and R-parity. Experiments so far failed to detect a non-vanishing size for any elementary particle. Page 1078

Moving entities are made of particles. To see how deep this result is, you can apply it to all those moving entities for which it is usually forgotten, such as ghosts, spirits, angels, nymphs, daemons, devils, gods, goddesses and souls. You can check yourself what happens when their particle nature is taken into account. Challenge 1349 e

From the existence of a minimum action, quantum theory deduces all its statements about particle motion. We go through the main ones.

▪ There is no rest for microscopic particles. All objects obey the indeterminacy principle, which states that the indeterminacies in position x and momentum p follow

$$\Delta x \Delta p \geqslant \hbar/2 \quad \text{with} \quad \hbar = 1.1 \cdot 10^{-34}\,\text{Js} \qquad (628)$$

and making rest an impossibility. The state of particles is defined by the same observables as in classical physics, with the difference that observables do not commute. Classical physics appears in the limit that the Planck constant $\hbar$ can effectively be set to zero.

▪ Quantum theory introduces a probabilistic element into motion. It results from the minimum action value through the interactions with the baths in the environment of any system.

▪ Large number of identical particles with the same momentum behave like waves. The so-called de Broglie wavelength λ is given by the momentum p of a single particle through

$$\lambda = \frac{h}{p} = \frac{2\pi\hbar}{p} \tag{629}$$

both in the case of matter and of radiation. This relation is the origin of the wave behaviour of light. The light particles are called photons; their observation is now standard practice. All waves interfere, refract and diffract. This applies to electrons, atoms, photons and molecules. All waves being made of particles, all waves can be seen, touched and moved. Light for example, can be 'seen' in photon-photon scattering, can be 'touched' using the Compton effect and it can be 'moved' by gravitational bending. Matter particles, such as molecules or atoms, can be seen, e.g. in electron microscopes, as well as touched and moved, e.g. with atomic force microscopes. The interference and diffraction of wave particles is observed daily in the electron microscope.

▪ Particles cannot be enclosed. Even though matter is impenetrable, quantum theory shows that tight boxes or insurmountable obstacles do not exist. Waiting long enough always allows to overcome boundaries, since there is a finite probability to overcome any obstacle. This process is called tunnelling when seen from the spatial point of view and is called decay when seen from the temporal point of view. Tunnelling explains the working of television tubes as well as radioactive decay.

▪ Particles are described by an angular momentum called spin, specifying their behaviour under rotations. Bosons have integer spin, fermions have half integer spin. An even number of bound fermions or any number of bound bosons yield a composite boson; an odd number of bound fermions or an infinite number of interacting bosons yield a low-energy fermion. Solids are impenetrable because of the fermion character of its electrons in the atoms.

▪ Identical particles are indistinguishable. Radiation is made of bosons, matter of fermions. Under exchange, fermions commute at space-like separations, whereas bosons anticommute. All other properties of quantum particles are the same as for classical particles, namely countability, interaction, mass, charge, angular momentum, energy, momentum, position, as well as impenetrability for matter and penetrability for radiation.

▪ In collisions, particles interact locally, through the exchange of other particles. When matter particles collide, they interact through the exchange of virtual bosons, i.e. off-shell bosons. Motion change is thus due to particle exchange. Exchange bosons of even spin mediate only attractive interactions. Exchange bosons of odd spin mediate repulsive interactions as well.

▪ The properties of collisions imply the existence of antiparticles, as regularly observed in experiments. Elementary fermions, in contrast to many elementary bosons, differ from their antiparticles; they can be created and annihilated only in pairs. Apart from neutrinos, elementary fermions have non-vanishing mass and move slower than light.

▪ Images, made of radiation, are described by the same properties as matter. Images can only be localized with a precision of the wavelength λ of the radiation producing it.

▪ The appearance of Planck's constant $\hbar$ implies that length scales exist in nature. Quantum theory introduces a fundamental jitter in every example of motion. Thus the infinitely small is eliminated. In this way, lower limits to structural dimensions and to many other measurable quantities appear. In particular, quantum theory shows that it is impossible that on the electrons in an atom small creatures live in the same way that humans live on the Earth circling the Sun. Quantum theory shows the impossibility of Lilliput.

▪ Clocks and meter bars have finite precision, due to the existence of a smallest action and due to their interactions with baths. On the other hand, all measurement apparatuses must contain baths, since otherwise they would not be able to record results.

Ref. 919 ▪ Quantum physics leaves no room for cold fusion, astrology, teleportation, telekinesis, supernatural phenomena, multiple universes, or faster than light phenomena – the EPR paradox notwithstanding.

Results of quantum field theory

Quantum field theory is that part of quantum theory that includes the process of transformation of particles into each other. The possibility of transformation results from the existence of a minimum action. Transformations have several important consequences. Page 665

▪ Quantum electrodynamics is the quantum field description of electromagnetism. Like all the other interactions, its Lagrangian is determined by the gauge group, the requirements of space-time (Poincaré) symmetry, permutation symmetry and renormalizability. The latter requirement follows from the continuity of space-time. Through the effects of virtual particles, QED describes decay, pair creation, vacuum energy, Unruh radiation for accelerating observers, the Casimir effect, i.e. the attraction of neutral conducting bodies, and the limit for the localization of particles. In fact, an object of mass m can be localized only within intervals of the Compton wavelength

$$\lambda_C = \frac{h}{mc} = \frac{2\pi\hbar}{mc} , \qquad (630)$$

where c is the speed of light. At the latest at these distances we must abandon the classical description and use quantum field theory. Quantum field theory introduces corrections to classical electrodynamics; among others, the nonlinearities thus appearing produce small departures from the superposition principle for electromagnetic fields, resulting in photon-photon scattering.

▪ Composite matter is separable because of the finite interaction energies of the constituents. Atoms are made of a nucleus made of quarks, and of electrons. They provide an effective minimal length scale to all everyday matter.

▪ Elementary particles have the same properties as either objects or images, except divisibility. The elementary fermions (objects) are: the six leptons electron, muon, tau, each with its corresponding neutrino, and the six quarks. The elementary bosons (images) are the photon, the eight gluons and the two weak interaction bosons. Page 1078

▪ Quantum chromodynamics, the field theory of the strong interactions, explains the masses of mesons and baryons through its descriptions as bound quark states. At fundamental scales, the strong interaction is mediated by the elementary gluons. At femtometer scales, the strong interaction effectively acts through the exchange of spin 0 pions, and is thus strongly attractive.

▪ The theory of electroweak interactions describes the unification of electromagnetism and weak interactions through the Higgs mechanism and the mixing matrix.

▪ Objects are composed of particles. Quantum field theory provides a complete list of Page 1078 the intrinsic properties which make up what is called an 'object' in everyday life, namely the same which characterize particles. All other properties of objects, such as shape, temperature, (everyday) colour, elasticity, density, magnetism, etc., are merely combinations of the properties from the particle properties. In particular, quantum theory specifies an object, like every system, as a part of nature interacting weakly and incoherently with its environment.

▪ Since quantum theory explains the origin of material properties, it also explains the origin of the properties of life. Quantum theory, especially the study of the electroweak and the strong forces, has allowed to give a common basis of concepts and descriptions to materials science, nuclear physics, chemistry, biology, medicine and to most of astronomy.

For example, the same concepts allow to answer questions such as why water is liquid at room temperature, why copper is red, why the rainbow is coloured, why the Sun and the stars continue to shine, why there are about 110 elements, where a tree takes the material to make its wood and why we are able to move our right hand at our own will.

▪ Matter objects are permanent because, in contrast to radiation, matter particles can only disappear when their antiparticles are present. It turns out that in our environment antimatter is almost completely absent, except for the cases of radioactivity and cosmic rays, where it appears in tiny amounts.

▪ The particle description of nature, e.g. particle number conservation, follows from the possibility to describe interactions perturbatively. This is possible only at low and medium energies. At extremely high energies the situation changes and non-perturbative effects come into play.

Is quantum theory magic?

Studying nature is like experiencing magic. Nature often looks different from what it is. During magic we are fooled – but only if we forget our own limitations. Once we start to see ourselves as part of the game, we start to understand the tricks. That is the fun of it. The same happens in physics.

▪ The world looks irreversible, even though it isn't. We never remember the future. We are fooled because we are macroscopic.

▪ The world looks decoherent, even though it isn't. We are fooled again because we are macroscopic.

▪ There are no clocks possible in nature. We are fooled because we are surrounded by a huge number of particles.

▪ Motion seems to disappear, even though it is eternal. We are fooled again, because our senses cannot experience the microscopic domain.

▪ The world seems dependent on the choice of the frame of reference, even though it is not. We are fooled because we are used to live on the surface of the Earth.

▪ Objects seem distinguishable, even though they are not. We are fooled because we live at low energies.

▪ Matter looks continuous, even though it isn't. We are fooled because of the limitations of our senses.

In short, our human condition permanently fools us. The answer to the title question is affirmative: quantum theory is magic. That is its main attraction.

The dangers of buying a can of beans

Another summary of our walk so far is given by the ultimate product warning, which according to certain well-informed lawyers should be printed on every cans of beans and on every product package. It shows in detail how deeply our human condition fools us.

Ref. 920

Warning: care should be taken when **looking** at this product:

- It emits heat radiation.
- Bright light has the effect to compress this product.

Warning: care should be taken when **touching** this product:

- Part of it could heat up while another part cools down, causing severe burns.

Warning: care should be taken when **handling** this product:

- This product consists of at least 99.999 999 999 999 % empty space.
- This product contains particles moving with speeds higher than one million kilometres per hour.
- Every kilogram of this product contains the same amount of energy as liberated by about one hundred nuclear bombs.*
- In case this product is brought in contact with antimatter, a catastrophic explosion will occur.
- In case this product is rotated, it will emit gravitational radiation.

Warning: care should be taken when **transporting** this product:

- The force needed depends on its velocity, as does its weight.
- This product will emit additional radiation when accelerated.
- This product attracts, with a force that increases with decreasing distance, every other object around, including its purchaser's kids.

Warning: care should be taken when **storing** this product:

- It is impossible to keep this product in a specific place and at rest at the same time.
- Except when stored underground at a depth of several kilometres, over time cosmic radiation will render this product radioactive.

* A standard nuclear warhead has an explosive yield of about 0.2 megatons (implied is the standard explosive trinitrotoluene or TNT), about thirteen times the yield of the Hiroshima bomb, which was 15 kiloton. A megaton is defined as 1 Pcal=4.2 PJ, even though TNT delivers about 5 % slightly less energy that this value. In other words, a megaton is the energy content of about 47 g of matter. That is less than a handful for most solids or liquids.

Ref. 921

- This product may disintegrate in the next 10^{35} years.
- It could cool down and lift itself into the air.
- Parts of this product are hidden in other dimensions.
- This product warps space and time in its vicinity, including the storage container.
- Even if stored in a closed container, this product is influenced and influences all other objects in the universe, including your parents in law.
- This product can disappear from its present location and reappear at any random place in the universe, including your neighbour's garage.

Warning: care should be taken when **travelling away from** this product:

- It will arrive at the expiration date before the purchaser does so.

Warning: care should be taken when **using** this product:

- Any use whatsoever will increase the entropy of the universe.
- The constituents of this product are exactly the same as those of any other object in the universe, including those of rotten fish.
- The use could be disturbed by the (possibly) forthcoming collapse of the universe.

The impression of a certain paranoid side to physics is purely coincidental.

The essence and the limits of quantum theory

We can summarize quantum physics with a simple statement: *quantum physics is the description of matter and radiation* without *the concept of infinitely small.* Matter and radiation are described by finite quantities. We had already eliminated the infinitely large in our exploration of relativity. On the other hand, some types of infinities remain. We had to retain the infinitely small in the description of space or time, and in topics related to them, such as renormalization. We did not manage to eliminate all infinities yet. We are thus not yet at the end of our quest. Surprisingly, we shall soon find out that a completely finite description of all of nature is equally impossible. To find out more, we focus on the path that remains to be followed.

What is unexplained by quantum theory and general relativity?

The material gathered in this second part of our mountain ascent, together with the earlier summary of general relativity, allows us to describe all observed phenomena connected to motion. Therefore, we are also able to provide a complete list of the *unexplained* properties of nature. Whenever we ask 'why?' about an observation and continue doing so after each answer, we arrive at one of the points listed in Table 68.

Table 68 *Everything* quantum field theory and general relativity do *not* explain; in other words, a list of *the only* experimental data and criteria available for tests of the unified description of motion

Observable	Property unexplained so far
Local quantities, from quantum theory	
α_{em}	the low energy value of the electromagnetic coupling constant

Observable	Property unexplained so far
α_w	the low energy value of the weak coupling constant
α_s	the low energy value of the strong coupling constant
m_q	the values of the 6 quark masses
m_l	the values of 3 lepton masses (or 6, if neutrinos have masses)
m_W	the values of the independent mass of the W vector boson
θ_W	the value of the Weinberg angle
$\beta_1, \beta_2, \beta_3$	three mixing angles (or 7, if neutrinos have masses)
θ_{CP}	the value of the CP parameter
θ_{st}	the value of the strong topological angle
3	the number of particle generations
$3+1$	the number of space and time dimensions
$0.5\,\mathrm{nJ/m^3}$	the value of the observed vacuum energy density or cosmological constant
Global quantities, from general relativity	
$1.2(1)\cdot 10^{26}\,\mathrm{m}$ (?)	the distance of the horizon, i.e. the 'size' of the universe (if it makes sense)
10^{82} (?)	the number of baryons in the universe, i.e. the average matter density in the universe (if it makes sense)
10^{92} (?)	the initial conditions for more than 10^{92} particle fields in the universe, including those at the origin of galaxies or stars (if or as long as they make sense)
Local structures, from quantum theory	
$S(n)$	the origin of particle identity, i.e. of permutation symmetry
Ren. group	the renormalization properties, i.e. the existence of point particles
SO(3,1)	the origin of Lorentz (or Poincaré) symmetry (i.e. of spin, position, energy, momentum)
C^*	the origin of the algebra of observables
Gauge group	the origin of gauge symmetry (and thus of charge, strangeness, beauty, etc.)
in particular, for the standard model:	
U(1)	the origin of the electromagnetic gauge group (i.e. of the quantization of electric charge, as well as the vanishing of magnetic charge)
SU(2)	the origin of weak interaction gauge group
SU(3)	the origin of strong interaction gauge group
Global structures, from general relativity	
maybe $\mathrm{R}\times \mathrm{S}^3$ (?)	the unknown topology of the universe (if it makes sense)

The table has several notable aspects.* First of all, neither quantum mechanics nor general relativity explain any property unexplained in the other field. The two theories do not

* Every now and then, researchers provide other lists of open questions. However, they all fall into the list above. The elucidation of dark matter and of dark energy, the details of the big bang, the modifications of general relativity by quantum theory, the mass of neutrinos, the quest for unknown elementary particles such as the inflaton field, magnetic monopoles or others, the functioning of cosmic high-energy particle

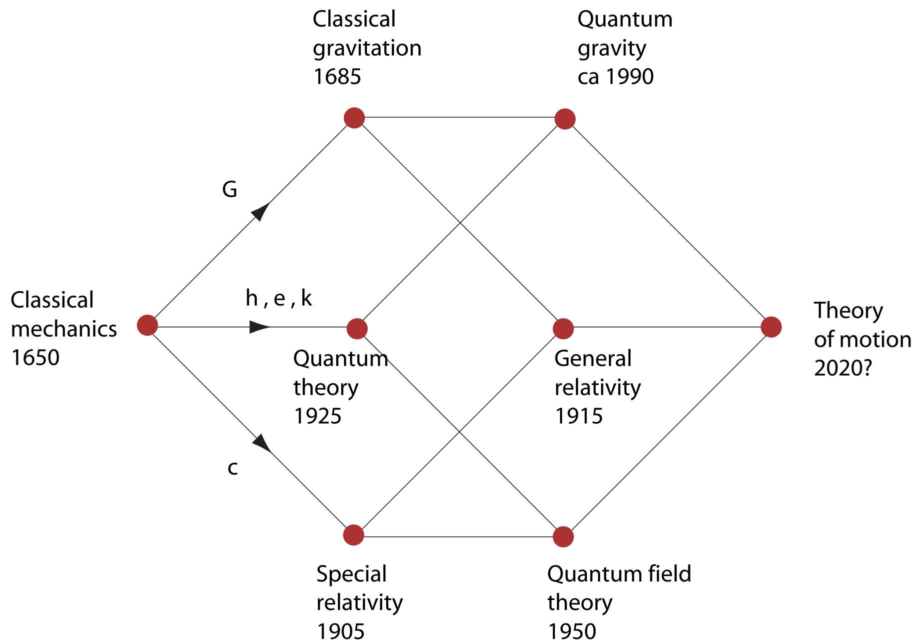


Figure 345 A simplified history of the description of motion in physics, by giving the limits to motion included in each description

help each other; the unexplained parts of both fields simply add up. Secondly, both in quantum theory and in general relativity, motion still remains the change of position with time. In short, in the first two parts of this walk we did not achieve our goal: we still do not understand motion. Our basic questions remain: What is time and space? What is mass? What is charge and what are the other properties of objects? What are fields? Why are all the electrons the same?

We also note that Table 68 lists extremely *different* concepts. That means that at this point of our walk there is *a lot* we do not understand. Finding the answers will not be easy, but will require effort.

On the other hand, the list of unexplained properties of nature is also *short*. The description of nature our adventure has produced so far is concise and precise. No discrepancies from experiments are known. In other words, we have a good description of motion *in practice*. Going further is unnecessary if we only want to improve measurement precision. Simplifying the above list is mainly important from the *conceptual* point of view. For this reason, the study of physics at university often stops at this point. However, even though we have *no* known discrepancies with experiments, we are *not* at the top of Motion Mountain, as Table 68 shows.

An even more suggestive summary of the progress and open issues of physics is shown in Figure 345. From one corner of a cube, representing Galilean physics, three edges – labelled G, c and $\hbar$, e, k – lead to classical gravity, special relativity and quantum theory.

accelerators, the stability or decay of protons, the origins of the heavy chemical elements, other interactions between matter and radiation or the possibility of higher spatial dimensions are questions that all fall into the table above.

Each constant implies a limit to motion; in the corresponding theory, this limit is taken into account. From these first level theories, corresponding parallel edges lead to general relativity, quantum field theory and quantum gravity, which take into account two of the limits.* From the second level theories, all edges lead to the last missing corner; that is the theory of motion. It takes onto account all limits found so far. Only this theory is a full or unified description of motion. The important point is that we already know all limits to motion. To arrive at the last point, no new experiments are necessary. No new knowledge is required. We only have to advance in the right direction, with careful thinking. Reaching the final theory of motion is the topic of the third part of our adventure.

Finally, we note from Table 68 that all progress we can expect about the foundations of motion will take place in two specific fields: cosmology and high energy physics.

How to delude oneself that one has reached the top of the Motion Mountain

Nowadays it is deemed chic to pretend that the adventure is over at the stage we have just reached.* The reasoning is as follows. If we change the values of the unexplained constants from Table 68 only ever so slightly, nature would look completely different from what it Ref. 923 does. The consequences have been studied in great detail; Table 69 gives an overview of the results.

Table 69 A selection of the consequences of changing the properties of nature

Observable	Change	Result
Local quantities, from quantum theory		
α_{em}	smaller:	only short lived, smaller and hotter stars; no Sun
	larger:	darker Sun, animals die of electromagnetic radiation, too much proton decay, no planets, no stellar explosions, no star formation, no galaxy formation
	+60%:	quarks decay into leptons
	+200%:	proton-proton repulsion makes nuclei impossible
α_{w}	−50%:	carbon nucleus unstable
	very weak:	no hydrogen, no p-p cycle in stars, no C-N-O cycle
	+2%:	no protons from quarks
	$G_F m_e^2 \not\approx \sqrt{G m_e^2}$:	either no or only helium in the universe

* Of course, the historical evolution of Figure 345 is a simplification. A less simplified diagram would use three different arrows for $\hbar$, e and k, making the figure a five-dimensional cube. However, not all of its Challenge 1350 ny corners would have dedicated theories (can you confirm this?), and moreover, the diagram would be much less appealing.

* Actually this attitude is not new. Only the arguments have changed. Maybe the greatest physicist ever, Ref. 922 James Clerk Maxwell, already fought against this attitude over a hundred years ago: 'The opinion seems to have got abroad that, in a few years, all great physical constants will have been approximately estimated, and that the only occupation which will be left to men of science will be to carry these measurements to another place of decimals. ... The history of science shows that even during that phase of her progress in which she devotes herself to improving the accuracy of the numerical measurement of quantities with which she has long been familiar, she is preparing the materials for the subjugation of new regions, which would have remained unknown if she had been contented with the rough methods of her early pioneers.'

Observable	Change	Result
	much larger:	no stellar explosions, faster stellar burning
α_s	−9%:	no deuteron, stars much less bright
	−1%:	no C resonance, no life
	+3.4%:	diproton stable, faster star burning
	much larger:	carbon unstable, heavy nuclei unstable, widespread leukaemia
n-p mass difference	larger:	neutron decays in proton inside nuclei; no elements
	smaller:	free neutron not unstable, all protons into neutrons during big bang; no elements
	smaller than m_e:	protons would capture electrons, no hydrogen atoms, star life much shorter
m_l changes:		
e-p mass ratio	much different:	no molecules
	much smaller:	no solids
3 generations	6-8:	only helium in nature
	>8:	no asymptotic freedom and confinement
Global quantities, from general relativity		
horizon size	much smaller:	no people
baryon number	very different:	no smoothness
	much higher:	no solar system
Initial condition changes:		
Moon mass	smaller:	small Earth magnetic field; too much cosmic radiation; widespread child skin cancer
Moon mass	larger:	large Earth magnetic field; too little cosmic radiation; no evolution into humans
Sun's mass	smaller:	too cold for the evolution of life
Sun's mass	larger:	Sun too short lived for the evolution of life
Jupiter mass	smaller:	too many comet impacts on Earth; extinction of animal life
Jupiter mass	larger:	too little comet impacts on Earth; no Moon; no dinosaur extinction
Oort cloud object number	smaller:	no comets; no irregular asteroids; no Moon; still dinosaurs
galaxy centre distance	smaller:	irregular planet motion; supernova dangers
initial cosmicspeed	+0.1%:	1000 times faster universe expansion
	−0.0001%:	universe recollapses after 10 000 years
vacuum energy density	change by 10^{-55}:	no flatness
3 + 1 dimensions	different:	no atoms, no planetary systems

Observable	Change	Result
Local structures, from quantum theory		
permutation symmetry	none:	no matter
Lorentz symmetry	none:	no communication possible
U(1)	different:	no Huygens principle, no way to *see* anything
SU(2)	different:	no radioactivity, no Sun, no life
SU(3)	different:	no stable quarks and nuclei
Global structures, from general relativity		
topology	other:	unknown; possibly correlated gamma ray bursts or star images at the antipodes

Some even speculate that the table can be condensed into a single sentence: if any para-
Ref. 924 meter in nature is changed, the universe would either have too many or too few black
Challenge 1351 r holes. However, the proof of this condensed summary is not complete yet.

Table 69, on the effects of changing nature, is overwhelming. Obviously, even the tiniest changes in the properties of nature are incompatible with our existence. What does this mean? Answering this question too rapidly is dangerous. Many fall into a common trap, namely to refuse admitting that the unexplained numbers and other properties need to be explained, i.e. deduced from more general principles. It is easier to throw in some irrational belief. The three most fashionable beliefs are that the universe is *created* or *designed*, that the universe is *designed for people*, or that the values are random, as our universe happens to be *one of many others*.

All these beliefs have in common that they have no factual basis, that they discourage further search and that they sell many books. Physicists call the issue of the first belief *fine tuning*, and usually, but not always, steer clear from the logical errors contained in the so
Page 635 common belief in 'creation' discussed earlier on. However, many physicists subscribe to the second belief, namely that the universe is designed for people, calling it the *anthropic principle*, even though we saw that it is indistinguishable both from the simian principle
Page 641 or from the simple request that statements be based on observations. In 2004, this belief has even become fashionable among older string theorists. The third belief, namely *multiple universes*, is a minority view, but also sells well.

Stopping our mountain ascent with a belief at the present point is not different from doing so directly at the beginning. Doing so used to be the case in societies which lacked the passion for rational investigation, and still is the case in circles which discourage the use of reason among their members. Looking for beliefs instead of looking for answers means to give up the ascent of Motion Mountain while pretending to have reached the top.

That is a pity. In our adventure, accepting the powerful message of Table 69 is one of the most awe-inspiring, touching and motivating moments. There is only one possible implication based on facts: the evidence implies that we are only a *tiny part* of the universe, but *linked* with all other aspects of it. Due to our small size and to all the connections with our environment, any imagined tiny change would make us disappear, like a water droplet is swept away by large wave. Our walk has repeatedly reminded us of this smallness and dependence, and overwhelmingly does so again at this point.

Having faced this powerful experience, everybody has to make up his own mind on whether to proceed with the adventure or not. Of course, there is no obligation to do so.

Challenge 1352 n

What awaits us?

The shortness of the list of unexplained aspects of nature means that *no additional experimental data* are available as check of the final description of nature. Everything we need to arrive at the final description of motion will probably be deduced from the experimental data given in this list, and from nothing else. In other words, future experiments will *not* help us – except if they change something in the list, as supersymmetry might do with the gauge groups or astronomical experiments with the topology issue.

This lack of new experimental data means that to continue the walk is a *conceptual* adventure only. We have to walk into storms raging near the top of Motion Mountain, keeping our eyes open, without any other guidance except our reason: this is not an adventure of action, but an adventure of the mind. And it is an incredible one, as we shall soon find out. To provide a feeling of what awaits us, we rephrase the remaining issues in six simple challenges.

What determines *colours*? In other words, what relations of nature fix the famous fine structure constant? Like the hero of Douglas Adams' books, physicists know the answer to the greatest of questions: it is 137.036. But they do not know the question.

What fixes the contents of a *teapot*? It is given by its size to the third power. But why are there only three dimensions? Why is the tea content limited in this way?

Was Democritus *right*? Our adventure has confirmed his statement up to this point; nature is indeed well described by the concepts of particles and of vacuum. At large scales, relativity has added a horizon, and at small scales, quantum theory added vacuum energy and pair creation. Nevertheless, both theories *assume* the existence of particles and the existence of space-time, and neither *predicts* them. Even worse, both theories completely fail to predict the existence of *any* of the properties either of space-time – such as its dimensionality – or of particles – such as their masses and other quantum numbers. A lot is missing.

Was Democritus *wrong*? It is often said that the standard model has only about twenty unknown parameters; this common mistake negates about 10^{93} initial conditions! To get an idea of the problem, we simply estimate the number N of possible states of all particles in the universe by

$$N = n\ v\ d\ p\ f \qquad (631)$$

where n is the number of particles, v is the number of variables (position, momentum, spin), d is the number of different values each of them can take (limited by the maximum of 61 decimal digits), p is the number of visible space-time points (about 10^{183}) and f is a factor expressing how many of all these initial conditions are actually independent of each other. We thus have the following number of possibilities

$$N = 10^{92} \cdot 8 \cdot 10^{61} \cdot 10^{183} \cdot f = 10^{336} \cdot f \qquad (632)$$

from which the 10^{93} actual initial conditions have to be explained. There is a small problem that we know nothing whatsoever about f. Its value could be 0, if all data were interdependent, or 1, if none were. Worse, above we noted that initial conditions cannot be defined for the universe at all; thus f should be undefined and not be a number at all! Whatever the case, we need to understand how all the visible particles get their 10^{93} states assigned from this range of options.

Page 760

Page 144

Were our efforts up to this point *in vain*? Quite at the beginning of our walk we noted that in classical physics, space and time are defined using matter, whereas matter is defined using space-time. Hundred years of general relativity and of quantum theory, including dozens of geniuses, have not solved this oldest paradox of all. The issue is still open at this point of our walk, as you might want to check by yourself.

Challenge 1353 e

The answers to these six questions define the top of Motion Mountain. Answering them means to know *everything* about motion. In summary, our quest for the unravelling of the essence of motion gets really interesting only from this point onwards!

> That is why Leucippus and Democritus, who say that the atoms move always in the void and the unlimited, must say what movement is, and in what their natural motion consists.
>
> Aristotle, *Treaty of the Heaven*

Ref. 925

Bibliography

919 An informative account of the world of psychokinesis and the paranormal is given by the famous professional magician James Randi, *Flim-flam!*, Prometheus Books, Buffalo 1987, as well as in several of his other books. See also the http://www.randi.org website. Cited on page 884.

920 This way to look at things goes back to the text by Susan Hewitt & Edward Subitzky, A call for more scientific truth in product warning labels, *Journal of Irreproducible Results* **36**, nr. 1, 1991. Cited on page 886.

921 J. Malik, The yields of the Hiroshima and Nagasaki nuclear explosions, Technical Report LA-8819, Los Alamos National Laboratory, September 1985. Cited on page 886.

922 James Clerk Maxwell, *Scientific Papers*, **2**, p. 244, October 1871. Cited on page 890.

923 A good introduction is C.J. Hogan, Why the universe is just so, *Reviews of Modern Physics* **72**, pp. 1149–1161, 2000. Most of the material of Table 69 is from the mighty book by John D. Barrow & Frank J. Tipler, *The Anthropic Cosmological Principle*, Oxford University Press, 1986. Discarding unrealistic options is also an interesting pastime. See for example the reasons why life can only be carbon-based, as explained in the essay by I. Asimov, The one and only, in his book *The Tragedy of the Moon*, Doubleday, Garden City, New York, 1973. Cited on page 890.

924 L. Smolin, The fate of black hole singularities and the parameters of the standard models of particle physics and cosmology, http://www.arxiv.org/abs/gr-qc/9404011. Cited on page 892.

925 ARISTOTLE, Treaty of the heaven, III, II, 300 b 8. See JEAN-PAUL DUMONT, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 392, 1991. Cited on page 894.

Bacteria, Flies and Knots

La première et la plus belle qualité de la nature est le mouvement qui l'agite sans cesse ; mais ce mouvement n'est qu'une suite perpétuelle de crimes ; ce n'est que par des crimes qu'elle le conserve.
Donatien de Sade, *Justine, ou les malheurs de la vertu.**

Wobbly entities, in particular jellyfish or amoebas, open up a fresh vision of the world of motion, if we allow to be led by the curiosity to study them in detail. They will yield many delightful insights about motion we have missed so far. In particular, wobbly entities yield surprising connections between shape change and motion which will be of great use in the last part of our mountain ascent. Instead of continuing to look at the smaller and smaller, we now take a look back, towards everyday motion and its mathematical description.

To enjoy this intermezzo, we change a dear habit. So far, we always described any general example of motion as composed of the motion of *point particles*. This worked well in classical physics, in general relativity and in quantum theory; we based the approach on the silent assumption that during motion, each point of a complex system can be followed separately. We will soon discover that this assumption is *not* realized at smallest scales. Therefore the most useful description of motion of *extended* bodies uses methods that do not require that body parts be followed one by one. We explore this issue in this intermezzo; doing so is a lot of fun in its own right.

If we imagine particles as extended entities – as we soon will have to – a particle moving through space is similar to a dolphin swimming through water or to a bee flying through air. Let us explore how these animals do this.

Bumble-bees and other miniature flying systems

When a butterfly passes by, as can happen to anybody ascending a mountain as long as flowers are present, we can stop a moment to appreciate a simple fact: a butterfly flies, and it is rather small. If you leave some cut fruit in the kitchen until it is rotten, we find the even smaller fruit flies. If you have ever tried to build small model aeroplanes, or if you even only compare them to paper aeroplanes (probably the smallest man-made flying thing you ever saw) you start to get a feeling for how well evolution optimized insects. Compared to paper planes, insects also have engines, flapping wings, sensors, navigation

* 'The primary and most beautiful of nature's qualities is motion, which agitates her at all times; but this motion is simply a perpetual consequence of crimes; she conserves it by means of crimes only.' Donatien Alphonse François de Sade (1740–1814) is the French writer from whom the term 'sadism' was deduced.

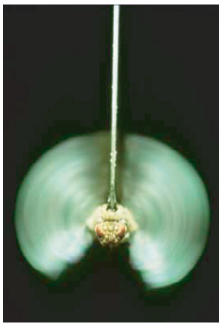


Figure 346 A flying fruit fly, tethered to a string

Figure 347 Vortices around a butterfly wing

systems, gyroscopic stabilizers, landing gear and of course all the features due to life, reproduction and metabolism, built into an incredibly small volume. Evolution really is an excellent engineering team. The most incredible flyers, such as the common house fly (*Musca domestica*), can change flying direction in only 30 ms, using the stabilizers that nature has built by reshaping the original second pair of wings. Human engineers are getting more and more interested in the technical solutions evolution has chosen and are

Ref. 927 trying to achieve the same miniaturization. The topic is extremely vast, so that we will pick out only a few examples.

How does an insect such as a fruit fly (*Drosophila melanogaster*) fly? The lift mg generated by a *fixed* wing follows the relation

$$mg = f\,A\,v^2\,\rho \qquad (633)$$

where A is the surface of the wing, v is the speed of the wing in the fluid of density ρ. The factor f is a pure number, usually with a value between 0.2 and 0.4, that depends on

Ref. 926 the angle of the wing and its shape; here we use the average value 0.3. For a Boeing 747, the surface is 511 m^2, the top speed is 250 m/s; at an altitude of 12 km the density of air is only a quarter of that on the ground, thus only 0.31 kg/m^3. We deduce (correctly) that

Challenge 1354 e a Boeing 747 has a mass of about 300 ton. For bumble-bees with a speed of 3 m/s and a wing surface of 1 cm^2, we get a lifted mass of about 35 mg, much less than the weight of the bee, namely about 1 g. In other words, a bee, like any other insect, cannot fly if it keeps its wings *fixed*. It could not fly with fixed wings even if it had propellers! Therefore, all insects must *move* their wings, in contrast to aeroplanes, not only to advance or to gain height, but also to simply remain airborne. Aeroplanes generate enough lift with fixed wings. Indeed, if you look at flying animals, you note that the larger they are, the less they need to move their wings.

Challenge 1355 n Can you deduce from equation (633) that birds or insects can fly but people cannot? The formula also (partly) explains why human powered aeroplanes must be so large.*

* The rest of the explanation requires some aerodynamics, which we will not study here. Aerodynamics shows that the power consumption, and thus the resistance of a wing with given mass and given cruise

But *how* do insects, small birds, flying fish or bats have to move their wings? This is a tricky question. In fact, the answer is just being uncovered by modern research. The main point is that insect wings move in a way to produce eddies at the front edge which in turn thrust the insect upwards. The aerodynamic studies of butterflies – shown in Figure 347 – and the studies of enlarged insect models moving in oil instead of in air are exploring the way insects make use of vortices. Researchers try to understand how vortices allow controlled flight at small dimensions. At the same time, more and more mechanical birds and model 'aeroplanes' that use flapping wings for their propulsion are being built around the world. The field is literally in full swing.* The aim is to reduce the size of flying machines. However, none of the human-built systems is yet small enough that it actually *requires* wing motion to fly, as is the case for insects.

Ref. 928

Ref. 929

Formula (633) also shows what is necessary for take-off and landing. The lift of wings decreases for smaller speeds. Thus both animals and aeroplanes increase their wing surface in these occasions. But even strongly flapping enlarged wings often are not sufficient at take-off. Many flying animals, such as swallows, therefore avoid landing completely. For flying animals which do take off from the ground, nature most commonly makes them hit the wings against each other, over their back, so that when the wings separate again, the vacuum between them provides the first lift. This method is used by insects and many birds, such as pheasants. As every hunter knows, pheasants make a loud 'clap' when they take off.

Both wing use and wing construction thus depend on size. There are four types of wings in nature. First of all, all large flying objects, such aeroplanes and large birds, fly using *fixed* wings, except during take-off and landing. Second, common size birds use *flapping* wings. These first two types of wings have a thickness of about 10 to 15 % of the wing depth. At smaller dimensions, a third wing type appears, as seen in dragonflies and other insects. At these scales, at Reynolds numbers of around 1000 and below, thin *membrane* wings are the most efficient. The *Reynolds number* measures the ratio between inertial and viscous effects in a fluid. It is defined as

Figure to be inserted

Figure 348 Two large wing types

$$R = \frac{l v \rho}{\eta} \tag{634}$$

where l is a typical length of the system, v the speed, ρ the density and η the dynamic *viscosity* of the fluid.* A Reynolds number much larger than one is typical for rapid air

speed, is inversely proportional to the square of the wingspan. Large wingspans with long slender wings are thus of advantage in (subsonic) flying, especially when energy needs to be conserved.

* The website http://www.aniprop.de presents a typical research approach and the sites http://ovirc.free.fr and http://www.ornithopter.org give introductions into the way to build such systems for hobbyists.

* The viscosity is the resistance to flow a fluid poses. It is defined by the force F necessary to move a layer of surface A with respect to a second, parallel one at distance d; in short, the (coefficient of) *dynamic viscosity* is defined as $\eta = d\,F/A\,v$. The unit is 1 kg/s m or 1 Pa s or 1 N s/m^2, once also called 10 P or 10 poise. In other words, given a horizontal tube, the viscosity determines how strong the pump needs to be to pump the fluid

flow and fast moving water. In fact, the Reynolds numbers specifies what is meant by a 'rapid' or 'fluid' flow on one hand, and a 'slow' or 'viscous' flow on the other. The first three wing types are all for rapid flows.

The fourth type of wings is found at the smallest possible dimensions, for insects smaller than one millimetre; their wings are not membranes at all. Typical are the cases of thrips and of parasitic wasps, which can be as small as 0.3 mm. All these small insects have wings which consist of a central *stalk* surrounded by hair. In fact, Figure 349 shows that some species of thrips have wings which look like miniature toilet brushes.

At even smaller dimensions, corresponding to Reynolds number below 10, nature does not use wings any more, though it still makes use of air transport. In principle, at the smallest Reynolds numbers gravity plays no role any more, and the process of flying merges with that of swimming. However, air currents are too strong compared with the speeds that such a tiny system could realize. No active navigation is then possible any more. At these small dimensions, which are important for the transport through air of spores and pollen, nature uses the air currents for passive transport, making use of special, but fixed shapes.

Figure 349 The wings of a few types of insects smaller than 1 mm (thrips, *Encarsia*, *Anagrus*, *Dicomorpha*)

We can summarize that active flying is only possible through shape change. Only two types of shape changes are possible: that of propellers (or turbines) and that of wings.* Engineers are studying with intensity how these shape changes have to take place in order to make flying most effective. Interestingly, the same challenge is posed by swimming. Ref. 930

Swimming

Swimming is a fascinating phenomenon. The Greeks argued that the ability of fish to swim is a proof that water is made of atoms. If atoms would not exist, a fish could not advance through it. Indeed, swimming is an activity that shows that matter cannot be continuous. Studying swimming can thus be quite enlightening. But how exactly do fish swim?

Whenever dolphins, jellyfish, submarines or humans *swim*, they take water with their fins, body, propellers, hands or feet and push it backwards. Due to momentum conservation they then move forward.** In short, people swim in the same way that fireworks

through the tube at a given speed. The viscosity of air 20°C is 1.8×10^{-5} kg/s m or 18 µPa s and increases with temperature. In contrast, the viscosity of liquids decreases with temperature. (Why?) The viscosity of water at 0°C is 1.8 mPa s, at 20°C it is 1.0 mPa s (or 1 cP), and at 40°C is 0.66 mPa s. Hydrogen has a viscosity smaller than 10 µPa s, whereas honey has 25 Pa s and pitch 30 MPa s. Challenge 1356 ny

Physicists also use a quantity ν called the *kinematic viscosity*. It is defined with the help of the mass density of the fluid as $\nu = \eta/\rho$ and is measured in m^2/s, once called 10^4 stokes. The kinematic viscosity of water at 20°C is $1\,mm^2/s$ (or 1 cSt). One of the smallest values is that of acetone, with $0.3\,mm^2/s$; a larger one is glycerine, with $2000\,mm^2/s$. Gases range between $3\,mm^2/s$ and $100\,mm^2/s$.

* The book by John Brackenbury, *Insects in Flight*, 1992. is a wonderful introduction into the biomechanics of insects, combining interesting science and beautiful photographs.

Page 69 ** Fish could use propellers, as the arguments against wheels we collected at the beginning of our walk do or rockets fly: by throwing matter behind them. Does all swimming work in this way? In particular, do *small* organisms advancing through the molecules of a liquid use the same method? No.

It turns out that small organisms such as bacteria do *not* have the capacity to propel or accelerate water against their surroundings. From far away, the swimming of microorganisms thus resembles the motion of particles through vacuum. Like microorganisms, also particles have nothing to throw behind them. Indeed, the water remains attached around a microorganism without ever moving away from it. Physically speaking, in these cases of swimming the kinetic energy of the water is negligible. In order to swim, unicellular beings thus need to use other effects. In fact, their only possibility is to change their body shape in controlled ways.

Figure 350 A swimming scallop (here from the genus *Chlamys*)

Let us go back to everyday scale for a moment. Swimming *scallops*, molluscs up to a few cm in size, can be used to clarify the difference between macroscopic and microscopic swimming. Scallops have a double shell connected by a hinge that they can open and close. If they close it *rapidly*, water is expelled and the mollusc is accelerated; the scallop then can glide for a while through the water. Then the scallop opens the shell again, this time *slowly*, and repeats the feat. When swimming, the larger scallops look like clockwork false teeth. If we reduce the size of the scallop by a thousand times to the size of single cells we get a simple result: such a tiny scallop *cannot* swim.

The origin of the lack of scalability of swimming methods is the changing ratio between inertial and dissipative effects at different scales. This ratio is measured by the Reynolds number. For the scallop the Reynolds number is about 100, which shows that when it swims, inertial effects are much more important than dissipative, viscous effects. For a bacterium the Reynolds number is much smaller than 1, so that inertial effects effectively play no role. There is no way to accelerate water away from a bacterial-sized scallop, and thus no way to glide. In fact one can even show the stronger result that no cell-sized being can move if the shape change is the same in the two halves of the motion (opening and closing). Such a shape change would simply make it move back and forward. Thus there is no way to move at cell dimensions with a method the scallop uses on centimetre scale; in fact the so-called *scallop theorem* states that no microscopic system can swim if it uses movable parts with only one degree of freedom. Ref. 931

Microorganisms thus need to use a more evolved, two-dimensional motion of their shape to be able to swim. Indeed, biologists found that all microorganisms use one of the following three swimming styles:

- 1. Microorganisms of compact shape of diameter between 20 μm and about 20 mm, use *cilia*. Cilia are hundreds of little hairs on the surface of the organism. The organisms move the cilia in waves wandering around their surface, and these surface waves make the body advance through the fluid. All children watch with wonder *Paramecium*, the unicel-

not apply for swimming. But propellers with blood supply would be a weak point in the construction, and thus in the defence of a fish.

lular animal they find under the microscope when they explore the water in which some grass has been left for a few hours. *Paramecium*, which is between 100 μm and 300 μm in size, as well as many plankton species* use cilia for its motion. The cilia and their motion are clearly visible in the microscope. A similar swimming method is even used by some large animals; you might have seen similar waves on the borders of certain ink fish; even the motion of the manta (partially) belongs into this class. Ciliate motion is an efficient way to change the shape of a body making use of two dimensions and thus avoiding the scallop theorem.

Ref. 932

▪ 2. Sperm and eukaryote microorganisms whose sizes are in the range between 1 μm and 50 μm swim using an (eukaryote) flagellum.* Flagella, Latin for 'small whips', work like flexible oars. Even though their motion sometimes appears to be just an oscillation, flagella get a kick only during one half of their motion, e.g. at every swing to the left. Flagella are indeed used by the cells like miniature oars. Some cells even twist their flagellum in a similar way that people rotate an arm. Some microorganisms, such as *Chlamydomonas*, even have two flagella which move in the same way as people move their legs when they perform the breast stroke. Most cells can also change the sense in which the flagellum is kicked, thus allowing them to move either forward or backward. Through their twisted oar motion, bacterial flagella avoid retracing the same path when going back and forward. As a result, the bacteria avoid the scallop theorem and manage to swim despite their small dimensions. The flexible oar motion they use is an example of a non-adiabatic mechanism; an important fraction of the energy is dissipated.

figure to be added

Figure 351 Ciliated and flagellate motion

Ref. 935
Ref. 939

▪ 3. The smallest swimming organisms, bacteria with sizes between 0.2 μm and 5 μm, swim using *bacterial* flagella. These flagella, also called prokaryote flagella, are different from the ones just mentioned. Bacterial flagella move like turning corkscrews. They are used by the famous *Escherichia coli* bacterium and by all bacteria of the genus *Salmonella*. This type of motion is one of the prominent exceptions to the non-existence of wheels in nature; we mentioned it in the beginning of our walk. Corkscrew motion is an example of an adiabatic mechanism.

Ref. 936
Page 69

A Coli bacterium typically has a handful of flagella, each about 30 nm thick and of corkscrew shape, with up to six turns; the turns have a 'wavelength' of 2.3 μm. Each flagellum is turned by a sophisticated rotation motor built into the cell, which the cell can control both in rotation direction and in angular velocity. For Coli bacteria, the range is between 0 and about 300 Hz.

Ref. 934

A turning flagellum does not propel a bacterium like a propeller; as mentioned, the velocities involved are much too small, the Reynolds number being only about 10^{-4}. At

* See the http://www.liv.ac.uk/ciliate/ website for an overview.

Ref. 933

* The largest sperm, of 6 cm length, are produced by the 1.5 mm sized *Drosophila bifurca* fly, a relative of the famous Drosophila melanogaster. Even when thinking about the theory of motion, it is impossible to avoid thinking about sex.

Figure 352 A well-known ability of cats

these dimensions and velocities, the effect is better described by a corkscrew turning in honey or in cork: a turning corkscrew produces a motion against the material around it, in the direction of the corkscrew axis. The flagellum moves the bacterium in the same way that a corkscrew moves the turning hand with respect to the cork.

Note that still smaller bacteria do not swim at all. Each bacterium faces a minimum swimming speed requirement: is must outpace diffusion in the liquid it lives in. Slow swimming capability makes no sense; numerous microorganisms therefore do not manage or do not try to swim at all. Some microorganisms are specialized to move along liquid–air interfaces. Others attach themselves to solid bodies they find in the liquid. Some of them are able to *move* along these solids. The amoeba is an example for a microorganism moving in this way. Also the smallest active motion mechanisms known, namely the motion of molecules in muscles and in cell membranes, work this way.

Ref. 937

Page 763

Let us summarize these observations in a different way. All known active motion, or self-propulsion, takes place in fluids – be it air or liquids. All active motion requires shape change. In order that shape change leads to motion, the environment, e.g. the water, must itself consist of moving components always pushing onto the swimming entity. The motion of the swimming entity can then be deduced from the particular shape change it performs. To test your intuition, you may try the following puzzle: is microscopic swimming possible in two spatial dimensions? In four?

Challenge 1357 ny

Falling cats and the theory of shape change

In the last decades, the theory of shape change has changed from a fashionable piece of research to a topic whose results are both appealing and useful. We have seen that shape change of a body in a fluid can lead to translation. But shape change can also lead to a *rotation* of the body. In particular, the theory of shape change is useful in explaining how falling cats manage to fall on their feet. Cats are not born with this ability; they have to learn it. But the feat remains fascinating. The great British physicist Michael Berry understood that this ability of cats can be described by an angular phase in a suitably defined shape space.

– CS – to be inserted – CS –

Page 86

In fact, cats confirm in three dimensions what we already knew for two dimensions: a deformable body can change its own orientation in space without outside help.

But shape change bears more surprises.

Turning a sphere inside out

> A text should be like a lady's dress; long enough to cover the subject, yet short enough to keep it interesting.

Continuing the theme of motion of wobbly entities, a famous example cannot be avoided. In 1957, the mathematician Stephen Smale proved that a sphere can be turned inside out. The discovery brought him the Fields medal in 1966, the highest prize for discoveries in mathematics. Mathematicians call his discovery the *eversion* of the sphere.

To understand the result, we need to describe more clearly the rules of mathematical eversion. First of all, it is assumed that the sphere is made of a thin membrane which has the ability to stretch and bend without limits. Secondly, the membrane is assumed to be able to *intersect* itself. Of course, such a ghostly material does not exist in everyday life; but in mathematics, it can be imagined. A third rule requires that the moves must be performed in such a way that the membrane is not punctured, ripped nor creased; in short, everything must happen *smoothly* (or differentiably, as mathematicians like to say).

Even though Smale proved that eversion is possible, the first way to actually perform it was discovered by the blind topologist Bernard Morin in 1961, based on ideas of Arnold Shapiro. After him, several additional methods have been discovered.

Several computer videos of sphere eversions are now available.* The most famous ones are Outside in, which shows an eversion due to William P. Thurston, and The Optiverse,

* Summaries of the videos can be seen at the http://www.geom.umn.edu/docs/outreach/oi website, which also has a good pedagogical introduction. Another simple eversion and explanation is given by Erik de Neve on the http://www.xs4all.nl/~alife/sphere1.htm website. It is even possible to run the movie software at home; see the http://www.cslub.uwaterloo.ca/~mjmcguff/eversion website. Figure 353 is from the http://new.math.uiuc.edu/optiverse website.

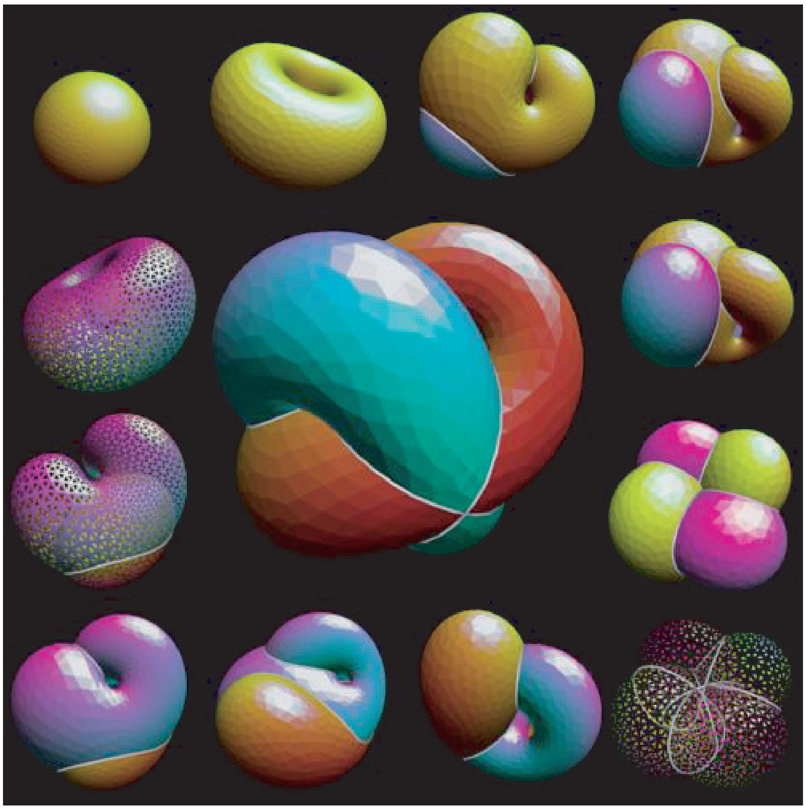


Figure 353 A way to turn a sphere inside out, with intermediate steps ordered clockwise

which shows the most efficient method known so far, discovered by a team led by John Sullivan and shown in Figure 353.

Why is sphere eversion of interest to physicists? If elementary particles were extended and at the same time were of spherical shape, eversion might be a symmetry of particles. To make you think, we mention the effects of eversion on the whole surrounding space, not only on the sphere itself. The final effect of eversion is the transformation

$$(x, y, z) \rightarrow \frac{(x, y, -z)\, R^2}{r^2} \tag{635}$$

where R is the radius of the sphere and r is the length of the coordinate vector (x, y, z), thus $r = \sqrt{x^2 + y^2 + z^2}$. Due to the minus sign in the z-coordinate, eversion is thus different from inversion, but not by too much. As we will find out shortly, a transformation similar to eversion, space-time duality, is a fundamental symmetry of nature.

Page 971

Knots, links and braids

Don't touch this, or I shall tie your fingers into knots!
(Surprisingly efficient child education technique.)

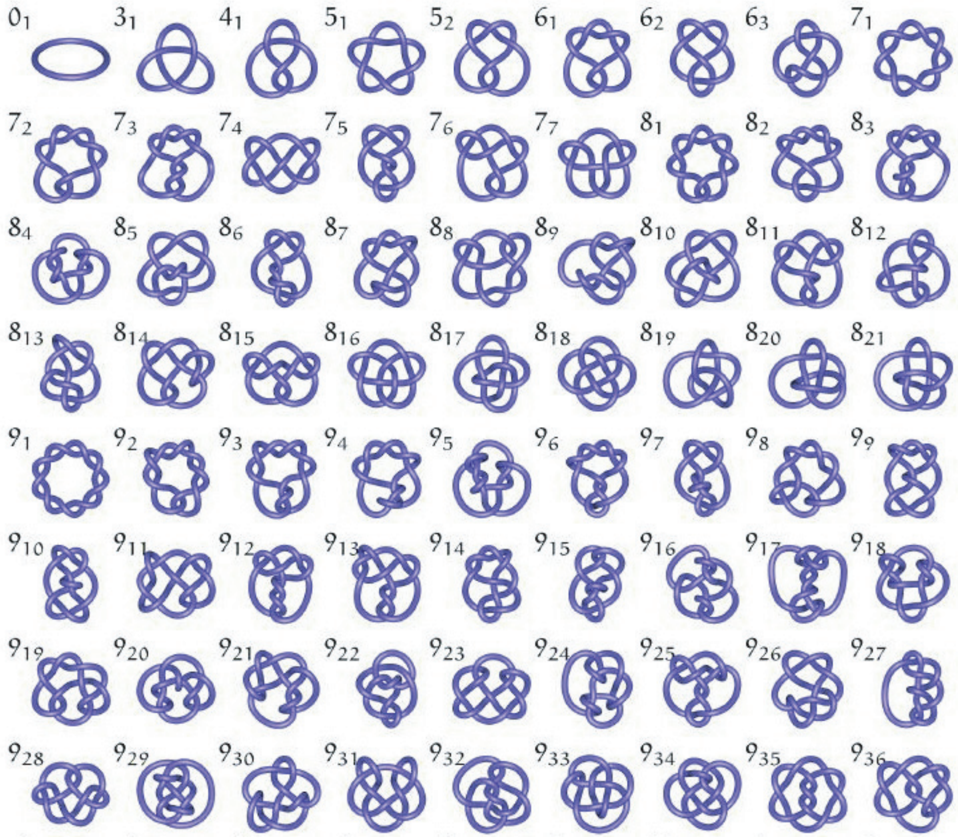


Figure 354 The knot diagrams for the simplest knots

Knots and their generalization are central to the study of wobbly entity motion. A *(mathematical) knot* is a closed piece of rubber string, i.e. a string whose ends have been glued together, which cannot be deformed into a circle or a simple loop. The simple loop is also called the *trivial knot*. If knots are ordered by their crossing numbers, as shown in Figure 354, the trivial knot (0_1) is followed by the trefoil knot (3_1) and by the figure-eight knot (4_1).

Knots are of importance in the context of this intermezzo as they visualize the limitations of the motion of wobbly entities. In addition, we will find other reasons to study knots later on. In this section, we just have a bit of fun.*

How do we describe such a knot through the telephone? Mathematicians have spent a lot of time to figure out smart ways to achieve it. The simplest way is to flatten the knot onto a plane and to list the position and the type (below or above) of the crossings.

Mathematicians are studying the *simplest* way to describe knots by the telephone. The task is not completely finished, but the end is in sight. Of course, the flat diagrams can be characterized by the minimal number of crossings. The knots in Figure 354 are ordered

* Pretty pictures and other information about knots can be found on the KnotPlot site, i.e. at the http://www.cs.ubc.ca/nest/imager/contributions/scharein/KnotPlot.html site.

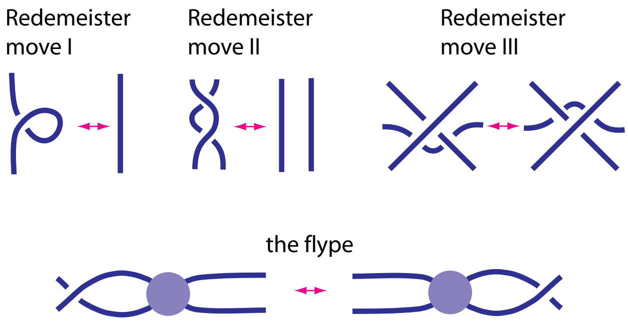


Figure 356 The Reidemeister moves and the flype

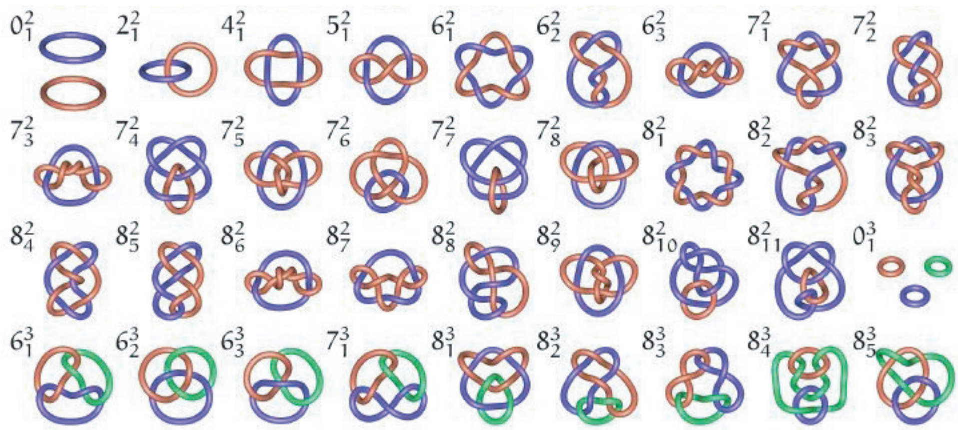


Figure 357 The diagrams for the simplest links with two and three components

in this way. There is 1 knot with zero, 1 with three and 1 with four crossings (not counting
Ref. 945 mirror knots); there are 2 knots with five and 3 with six crossings, 7 knots with seven, 21 knots with eight, 41 with nine, 165 with ten, 552 with eleven, 2176 with twelve, 9988 with thirteen, 46 972 with fourteen, 253 293 with fifteen and 1 388 705 knots with sixteen crossings.

Mathematicians do not talk about 'telephone messages', they talk about invariants, i.e. about quantities that do not depend on the precise shape of the knot. At present, the best description of knots is a polynomial invariant based on a discovery by Vaughan Jones in 1984. However, though the polynomial allows to uniquely describe most simple knots, it fails to do so for more complex ones. But the Jones polynomial finally allowed to prove that a diagram which is alternating and eliminates nugatory crossings (i.e. if it is 'reduced') is indeed one which has minimal number of crossings. The polynomial also allows to show that any two reduced alternating diagrams are related by a sequence of flypes.

Together with the search for invariants, the tabulation of knots is a modern mathematical sport. In 1949, Schubert proved that every knot can be decomposed in a unique way
Ref. 945 as sum of prime knots. Knots thus behave similarly to integers.

The mirror image of a knot usually, but not always, is different from the original. If you want a challenge, try to show that the trefoil knot, the knot with three crossings, is different from its mirror image. The first proof was by Max Dehn in 1914.

Antiknots do not exist. An *antiknot* would be a knot on a rope that cancels out the corresponding knot when the two are made to meet along the rope. It is easy to prove that this is impossible. We take an infinite sequence of knots and antiknots on a string, $K - K + K - K + K - K$.... On one hand, we could make them disappear in this way $K - K + K - K + K - K... = (K - K) + (K - K) + (K - K)... = 0$. On the other hand, we could do the same thing using $K - K + K - K + K - K... = K(-K + K) + (-K + K) + (-K + K)... = K$. The only knot K with an antiknot is thus the unknot $K = 0$.*

Ref. 947

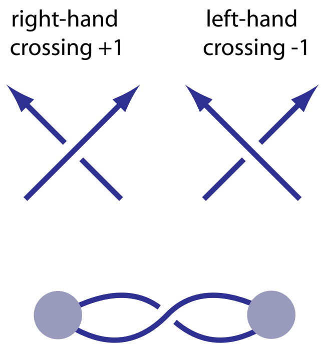


Figure 355 Crossing types in knots

– CS – Several topics to be included – CS –

Since knots are stable in time, a knotted line in three dimensions is equivalent to a knotted surface in space-time. When thinking in higher dimensions, we need to be careful. Every knot (or knotted line) can be untied in four or more dimensions; however, there is no surface *embedded* in four dimensions which has as $t = 0$ slice a knot, and as $t = 1$ slice the circle. Such a surface embedding needs at least five dimensions.

In higher dimensions, knots are possible only n-spheres are tied instead of circles; for example, as just said, 2-spheres can be tied into knots in 4 dimensions, 3-spheres in 5 dimensions and so forth.

Mathematicians also study more elaborate structures. *Links* are the generalization of knots to several closed strands. *Braids* are the generalization of links to *open* strands. Braids are especially interesting, as they form a group; can you state what the group operation is?

Challenge 1359 e

Knots in nature and on paper

Knots do not play a role only in shoe laces and in sailing boats.

- Proteins, the molecules that make up many cell structures, are chains of aminoacids. It seems that very few proteins are knotted, and that most of these form trefoil knots. However, a figure-eight knotted protein has been discovered in 2000 by William Taylor. (Ref. 949)
- Knots form also in other polymers. They seem to play a role in the formation of radicals in carbohydrates. Research on knots in polymers is presently in full swing.
- A famous type of eel, the *knot fish Myxine glutinosa*, also called hagfish or slime eel, is able to make a knot in his body and move this knot from head to tail. It uses this motion to cover its body with a slime that prevents predators from grabbing it; it also uses this motion to escape the grip of predators, to get rid of the slime after the danger is over, and (Ref. 946)

* This proof does *not* work when performed with numbers; we would be able to deduce $1 = 0$ by setting K=1. Why is this proof valid with knots but not with numbers?

Challenge 1358 n

Figure 358 A hagfish tied into a knot

Figure to be included

Figure 359 How to simulate order for long ropes

to push against a prey it is biting in order to extract a piece of meat. All studied knot fish form only left handed trefoil knots, by the way; this is another example of chirality in nature.

▪ One of the most incredible discoveries of recent years is related to knots in DNA molecules. The DNA *molecules* inside cell nuclei can be hundreds of millions of base pairs long; they regularly need to be packed and unpacked. When this is done, often the same happens as when a long piece of rope or a long cable is taken out of a closet.

It is well known that you can roll up a rope and put it into a closet in such a way that it looks orderly stored, but when it is pulled out at one end, a large number of knots is suddenly found. Figure 359 shows how to achieve this.

To make a long story short, this also happens to nature when it unpacks DNA in cell nuclei. Life requires that DNA molecules move inside the cell nucleus without hindrance. So what does nature do? Nature takes a simpler approach: when there are unwanted crossings, it cuts the DNA, moves it over and puts the ends together again. In cell nuclei, there are special enzymes, the so-called topoisomerases, which perform this process. The details of this fascinating process are still object of modern research.

▪ The great mathematician Carl-Friedrich Gauß was the first person to ask what would happen when an electrical current I flows along a wire A linked with a wire B. He discovered a beautiful result by calculating the effect of the magnetic field of one wire onto the other. Gauss found the expression Ref. 950

$$\frac{1}{4\pi I}\int_A \mathrm{d}\mathbf{x_A}\cdot\mathbf{B_B} = \frac{1}{4\pi}\int_A \mathrm{d}\mathbf{x_A}\cdot\int_B \mathrm{d}\mathbf{x_B}\times\frac{(\mathbf{x_A}-\mathbf{x_B})}{|\mathbf{x_A}-\mathbf{x_B}|^3} = n\ , \qquad (636)$$

where the integrals are performed along the wires. Gauss found that the number n does not depend on the precise shape of the wires, but only on the way they are linked. Deforming the wires does not change it. Mathematicians call such a number a *topological invariant*. In short, Gauss discovered a physical method to calculate a mathematical invariant for links; the research race to do the same for other invariants, also for knots and braids, is still going on today.

In the 1980s, Edward Witten was able to generalize this approach to include the nuclear interactions, and to define more elaborate knot invariants, a discovery that brought him the Fields medal.

▪ Knots are also of importance at Planck scales, the smallest dimensions possible in nature. We will soon explore how knots and the structure of elementary particles are related.

▪ Knots appear rarely in nature. For example, tree roots do not seem to grow many knots during the lifetime of a plant. How do plants avoid this? In other words, why are there no knotted bananas in nature? Challenge 1360 r

▪ If we move along the knot and count the crossings where we stay above and subtract the number of crossings where we pass below, we get a number called the *writhe* of the knot. It is not an invariant, but usually a tool in building them. The writhe is not necessarily invariant under one of the three Reidemeister moves. Can you see which one? Challenge 1361 ny However, the writhe is invariant under flypes.

Clouds

Clouds are another important class of extended entities. The lack of a definite boundary makes them even more fascinating than amoebas, bacteria or falling cats. We can observe the varieties of clouds from an aeroplane. We also have encountered clouds as the basic structure determining the size of atoms. Comparing these two and other types of clouds teaches us several interesting things about nature.

Galaxies are clouds of stars; stars are clouds of plasma; the atmosphere is a gas cloud. Page 551 Obviously, the common cumulus or cumulonimbus in the sky are vapour and water droplet clouds. Clouds of all types can be described by a shape and a size, even though in theory they have no bound. An effective shape can be defined by that region in which the cloud density is only, say, 1 % of the maximum density; slightly different procedures can also be used. All clouds are described by probability densities of the components making up the cloud. All clouds show conservation of their number of constituents.

Whenever we see a cloud, we can ask why it does not collapse. Every cloud is an aggregate. Page 168 All aggregates are kept from collapse in only three ways: through rotation, through pressure or through the Pauli principle, i.e. the quantum of action. Galaxies are kept from collapsing by rotation. Most stars and the atmosphere are kept from collapsing by gas pressure. Neutron stars, the Earth, atomic nuclei, protons or the electron clouds of atoms are kept apart by the quantum of action.

A rain cloud can contain several thousand tons of water; can you explain what keeps it afloat, and what else keeps it from continuously diffusing into a thinner and thinner structure? Challenge 1362 ny

Two rain clouds can merge. So can two atomic electron clouds. But only atomic clouds are able to cross each other. We remember that a normal atom can be inside a Rydberg atom and leave it again without change. Rain clouds, stars, galaxies or other macroscopic clouds cannot cross each other. When their paths cross, they can only merge or be ripped into pieces. Due to this lack of crossing ability, it is in fact easier to count atomic clouds than macroscopic clouds. In the macroscopic case, there is no real way to define a 'single' cloud in an accurate way. If we aim for full precision, we are unable to claim that there is more than one rain cloud, as there is no clear-cut boundary between them. Electronic clouds are different. True, in a piece of solid matter we can argue that there is only a single electronic cloud throughout the object; however, when the object is divided, the cloud is divided in a way that makes the original atomic clouds reappear. We thus can speak of 'single' electronic clouds.

Let us explore the limits of the topic. In our definition of the term 'cloud' we assumed that space and time are continuous. We also assumed that the cloud constituents were localized entities. This does not have to be the case.

Fluid space-time

So far, we have looked at the motion of wobbly entities in continuous space-time. But that is an unnecessary restriction. Looking at space-time itself in this way is also interesting. Ref. 951 The most intriguing approach was published in 1995 by Ted Jacobson. He explored what happens if space-time, instead of assumed to be continuous, is assumed to be the statistical average of numerous components moving in a disordered fashion.

The standard description of general relativity describes space-time as an entity similar Page 350 to a flexible mattress. Jacobson studied what happens if the mattress is assumed to be made of a liquid. A liquid is a collection of (undefined) components moving randomly and described by a temperature varying from place to place. He thus explored what happens if spacetime is made of fluctuating entities.

Page 800 Jacobson started from the Fulling–Davies–Unruh effect and assumed that the local temperature is given by the same multiple of the local gravitational acceleration. He also used the proportionality – correct on horizons – between area and entropy. Since the energy flowing through a horizon can be called heat, one can thus translate the expression $\delta Q = T\delta S$ into the expression $\delta E = a\delta A(c^2/4G)$, which describes the behaviour of Page 989 space-time at horizons. As we have seen, this expression is fully equivalent to general relativity.

In other words, imagining space-time as a liquid is a powerful analogy that allows to deduce general relativity. Does this mean that space-time actually *is* similar to a liquid? So far, the analogy is not sufficient to answer the question. In fact, just to confuse the reader a bit more, there is an old argument for the opposite statement.

Solid space-time

The main reason to try to model empty space as a solid is a famous property of the motion of dislocations. To understand it, a few concepts need to be introduced. *Dislocations* are one-dimensional construction faults in crystals, as shown in Figure 360. A general dislocation is a mixture of the two pure dislocation types: *edge dislocations* and *screw dislocations*. Both are shown in Figure 360. If one studies how the involved atoms can rearrange themselves, one finds that edge dislocations can only move perpendicularly to the added plane. In contrast, screw dislocations can move in all directions.* An important case of general, mixed dislocations, i.e. of mixtures of edge and screw dislocations, are closed *dislocation rings*. On such a dislocation ring, the degree of mixture changes continuously from place to place.

A dislocation is described by its strength and by its effective size; they are shown, respectively, in red and blue in Figure 360. The *strength* of a dislocation is measured by

* See the http://uet.edu.pk/dmems/edge_dislocation.htm, http://uet.edu.pk/dmems/screw_dislocation.htm and http://uet.edu.pk/dmems/mixed_dislocation.htm web pages for seeing a moving dislocation.

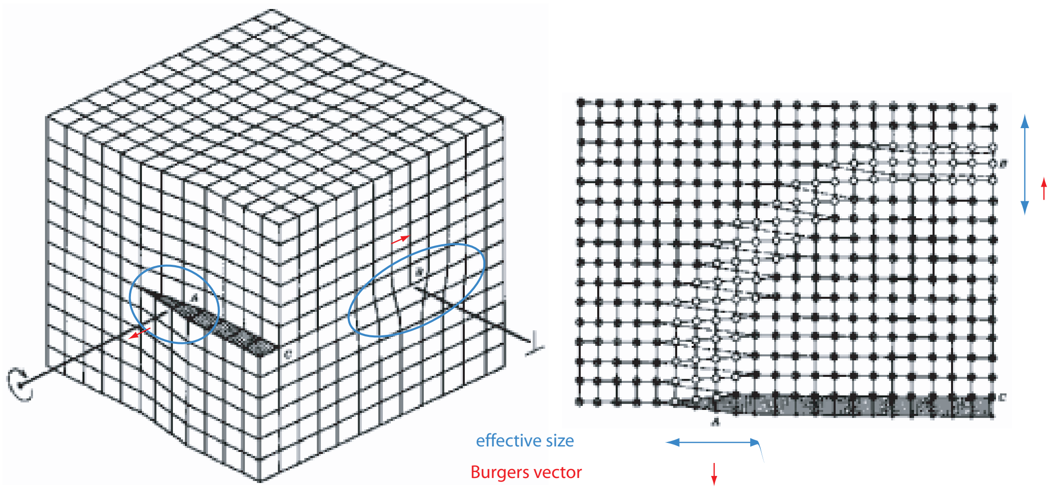


Figure 360 The two pure dislocation types: edge and screw dislocations

the so-called *Burgers vector*; it measures the misfits of the crystal around the dislocation. More precisely, the Burgers vector specifies by how much a section of perfect crystal needs to be displaced, after it has been cut open, to produce the dislocation. Obviously, the strength of a dislocation is quantized in multiples of a minimal Burgers vector. In fact, dislocations with large Burgers vectors can be seen as composed of dislocations of minimal Burgers vector.

The size or *width* of a dislocation is measured by an *effective width* w. Also the width is a multiple of the lattice vector. The width measures the size of the deformed region of the crystal around the dislocation. Obviously, the size of the dislocation depends on the elastic properties of the crystal, can take continuous values and is direction-dependent. The width is thus related to the energy content of a dislocation.

A general dislocation can move, though only in directions which are both perpendicular to its own orientation and to its Burgers vector. Let us study this motion in more detail. We call c the speed of sound in a pure (cubic) crystal. As Frenkel and Kontorowa
Ref. 952 found in 1938 it turns out that when a screw dislocation moves with velocity v, its width w changes as

$$w = \frac{w_0}{\sqrt{1 - v^2/c^2}} . \quad (637)$$

In addition, the energy of the moving dislocation obeys

$$E = \frac{E_0}{\sqrt{1 - v^2/c^2}} . \quad (638)$$

A screw dislocation thus cannot move faster than the speed of sound in a crystal and its width shows a speed-dependent contraction. (Edge dislocations have similar, but more complex behaviour.) The motion of screw dislocations in solids is described by the same effects and formulae that describe the motion of bodies in special relativity; the speed of

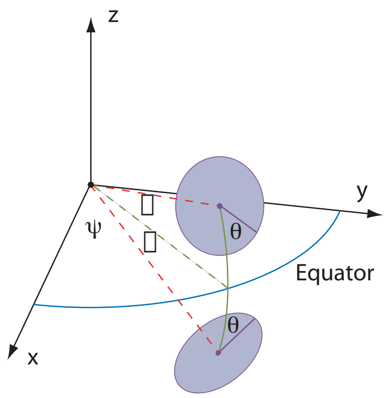


Figure 361 Swimming on a curved surface using two discs

sound is the limit speed for dislocations in the same way that the speed of light is the limit speed for objects.

Does this mean that elementary particles are dislocations of space or even of space-time, maybe even dislocation rings? The speculation is appealing, even though it supposes that space-time is a solid, and thus contradicts the model of space or space-time as a fluid. Page 932 Worse, we will soon encounter good reasons to reject modelling space-time as a lattice; Challenge 1363 n maybe you can find a few ones already by yourself. Still, expressions (637) and (638) for dislocations continue to fascinate. For the time being, we do not study them further.

Swimming in curved space

There is an additional reason to see space as a liquid. It is possible to *swim* through Ref. 953 empty space. This discovery was published in 2003 by Jack Wisdom. He found that cyclic changes in the shape of a body can lead to net translation, a rotation of the body, or both.

Swimming in space-time does not happen at high Reynolds numbers. That would imply that a system would be able to throw empty space behind it, and to propel itself forward as a result. No such effects have ever been found. However, Jack Wisdom found a way to swim that corresponds to low Reynolds numbers, where swimming results of simple shape change.

There is a simple system that shows the main idea. We know from Galilean physics that on a frictionless surface it is impossible to move, but that it is possible to turn oneself. This is true only for a flat surface. On a curved surface, one can use the ability to turn and translate it into motion.

Take to massive discs that lie on the surface of a frictionless, spherical planet, as shown in Figure 361. Consider the following four steps: The disc separation φ is increased by the angle $\Delta\varphi$, then the discs are rotated oppositely about their centres by the angle $\Delta\theta$, their separation is decreased by $-\Delta\varphi$, and they are rotated back by $-\Delta\theta$. Due to to the

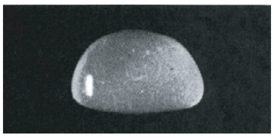


Figure 362 A large raindrop falling downwards

Figure 363 Is this possible?

Challenge 1364 ny conservation of angular momentum, the two-disc system changes its longitude $\Delta\psi$ as

$$\Delta\psi = \frac{1}{2}\gamma^2 \Delta\theta \Delta\varphi \,, \tag{639}$$

where γ is the angular radius of the discs. This cycle can be repeated over and over. The cycle it allows a body on the surface of the Earth, to swim along the surface. However, for a body of meter size, the motion for each swimming cycle is only around 10^{-27} m.

Wisdom showed that the mechanism also works in curved space-time. The mechanism thus allows a falling body to swim away from the path of free fall. Unfortunately, the achievable distances for everyday objects are negligible. Nevertheless, the effect exists.

At this point, we are thoroughly confused. Space-time seems to be solid and liquid at the same time. Despite this contrast, the situation gives the impression that extended, wobbly and fluctuating entities might lead us towards a better understanding of the structure of space and time. That exploration is left for the third and last part of our adventure.

Curiosities and fun challenges

> Any pair of shoes proves that we live on the inside of a sphere. Their soles are worn out at the ends, and hardly at all in between.
>
> Anonymous

The topic of wobbly entities is full of fascinating details. Here are a few.

Challenge 1365 n ▪ What is the shape of raindrops? Try to picture it. However, use your reason, not your prejudice! By the way, it turns out that there is a maximum size for raindrops, with a value of about 4 mm. The shape of such a large raindrop is shown in Figure 362. Can you imagine where the limit comes from? (Ref. 954)

For comparison, the drops in clouds, fog or mist are in the range of 1 to 100 μm, with a peak at 10 to 15 μm. In those cases when all droplets are of similar size one and when light is scattered only once by the droplets, one can observe coronae, glories or fogbows.

Challenge 1366 n ▪ What is the entity shown in Figure 363 – a knot, a braid or a link?

Challenge 1367 d ▪ Can you find a way to classify tie knots?

Challenge 1368 n ▪ Are you able to find a way to classify the way shoe laces can be threaded?

Outlook

We have studied one example of motion of extended bodies already earlier on: solitons. We can thus sum up the possible motions of extended entities in four key themes. We first studied solitons and interpenetration, then knots and their rearrangement, continued with duality and eversion and finally explored clouds and extension. The sum of it all seems to be half liquid and half solid.

The motion of wobbly bodies probably is the most neglected topic in all textbooks on motion. Research is progressing at full speed; it is expected that many beautiful analogies will be discovered in the near future. For example, in this intermezzo we have not described any good analogy for the motion of light; similarly, including quantum theory into the description of wobbly bodies' motion remains a fascinating issue for anybody aiming to publish in a new field. Challenge 1369 d

The ideas introduced in this intermezzo were sufficient to prepare us for the third part of our ascent of Motion Mountain. We can now tackle the final part of our adventure.

Bibliography

926 Henk Tennekes, *De wetten van de vliegkunst – over stijgen, dalen, vliegen en zweven*, Aramith Uitgevers, 1992. This clear and interesting text is also available in English. Cited on page 897.

927 In 1997, the smallest human-made flying object was the helicopter built by a group of the Institut für Mikrotechnik in Mainz, in Germany. A picture is available at their web page, to be found at http://www.imm-mainz.de/English/billboard/f_hubi.html. The helicopter is 24 mm long, weighs 400 mg and flies (though not freely) using two built-in electric motors driving two rotors, running at between 40 000 and 100 000 revolutions per minute. See also the helicopter from Stanford University at http://www-rpl.stanford.edu/RPL/htmls/mesoscopic/mesicopter/mesicopter.html, with an explanation of its battery problems. Cited on page 897.

928 The most recent computational models of lift still describe only two-dimensional wing motion, e.g. Z.J. Wang, Two dimensional mechanism for insect hovering, *Physical Review Letters 85* pp. 2216–2219, 2000. A first example of a mechanical bird has been constructed by Wolfgang Send; it can be studied on the http://www.aniprop.de website. See also W. Send, Physik des Fliegens, *Physikalische Blätter* 57, pp. 51–58, June 2001. Cited on page 898.

929 R.B. Srygley & A.L.R. Thomas, Unconventional lift-generating mechanisms in free-flying butterflies, *Nature* **420**, pp. 660–664, 2002. Cited on page 898.

930 The simulation of insect flight using enlarged wing models flapping in oil instead of air is described for example in http://www.dickinson.caltech.edu/research_robofly.html. The higher viscosity of oil allows to achieve the same Reynolds number with larger sizes and lower frequencies than in air. Cited on page 899.

931 E. Purcell, Life at low Reynolds number, *American Journal of Physics* **45**, p. 3, 1977. Cited on page 900.

932 Most bacteria are flattened, ellipsoidal sacks kept in shape by the membrane enclosing the cytoplasma. But there are exceptions; in salt water, quadratic and triangular bacteria have

been found. More is told in the corresponding section in the interesting book by Bernard Dixon, *Power Unseen – How Microbes Rule the World*, W.H. Freeman, New York, 1994. Cited on page 901.

933 S. Pitnick, G. Spicer & T.A. Markow, How long is a giant sperm?, *Nature* **375**, p. 109, 1995. Cited on page 901.

934 For an overview of the construction and the motion of coli bacteria, see H.C. Berg, Motile behavior of bacteria, *Physics Today* **53**, pp. 24–29, January 2000. Cited on page 901.

935 M. Kawamura, A. Sugamoto & S. Nojiri, Swimming of microorganisms viewed from string and membrane theories, *Modern Journal of Physics Letters A* **9**, pp. 1159–1174, 1994. Also available as http://www.arxiv.org/abs/hep-th/9312200. Cited on page 901.

936 They are also called prokaryote flagella. See for example S.C. Schuster & S. Khan, The bacterial flagellar motor, *Annual Review of Biophysics and Biomolecular Structure* **23**, pp. 509–539, 1994, or S.R. Caplan & M. Kara-Ivanov, The bacterial flagellar motor, *International Review of Cytology* **147**, pp. 97–164, 1993. See also the information on the topic that can be found on the website http://www.id.ucsb.edu:16080/fscf/library/origins/graphics-captions/flagellum.html. Cited on page 901.

937 This is from the book by David Dusenbery, *Life at a Small Scale*, Scientific American Library, 1996. Cited on page 902.

938 A. Shapere & F. Wilczek, Gauge kinematics of deformable bodies, *American Journal of Physics* **57**, pp. 514–518, 1989, A. Shapere & F. Wilczek, Geometry of self-propulsion at low Reynolds number, *Journal of Fluid Mechanics* **198**, pp. 557–585, 1989, A. Shapere & F. Wilczek, Efficiencies of self-propulsion at low Reynolds number, *Journal of Fluid Mechanics* **198**, pp. 587–599, 1989. See also R. Montgomery, Gauge theory of the falling cat, *Field Institute Communications* **1**, pp. 75–111, 1993. No citations.

939 W. Nutsch & U. Rüffer, Die Orientierung freibeweglicher Organismen zum Licht, dargestellt am Beispiel des Flagellaten Chlamydomonas reinhardtii, *Naturwissenschaften* **81**, pp. 164–174, 1994. Cited on page 901.

940 A lot here is from Louis H. Kauffman, *Knots and Physics*, World Scientific, second edition, 1994, which gives a good, simple, and visual introduction to the mathematics of knots and its main applications to physics. No citations.

941 S. Smale, A classification of immersions of the two-sphere, *Transactions of the American Mathematical Society* **90**, pp. 281–290, 1958. No citations.

942 G.K. Francis & B. Morin, Arnold Shapiro's Eversion of the Sphere, *Mathematical Intelligencer* pp. 200–203, 1979. See also reference Ref. 955. No citations.

943 Silvio Levy, Delle Maxwell & Tamara Munzner, *Making Waves – a Guide to the Ideas Behind Outside In*, Peters, 1995. No citations.

944 B. Morin & J.-P. Petit, Le retournement de la sphere, *Pour la Science* **15**, pp. 34–41, 1979. See also the clear article by A. Phillips, Turning a surface inside out, *Scientific American* pp. 112–120, May 1966. No citations.

945 A good introduction to knots is the paper by J. Hoste, M. Thistlethwaite & J. Weeks, The first 1,701,936 knots, *The Mathematical Intelligencer* **20**, pp. 33–47, 1998. Cited on page 906.

946 Alexei Sossinsky, *Nœuds – histoire d'une théorie mathématique*, Editions du Seuil, Paris, 1999. D. Jensen, Le poisson noué, *Pour la science*, dossier hors série, pp. 14–15, April 1997. Cited on page 907.

947 I. Stewart, *Game, Set and Math*, Penguin Books, 1989, pp. 58–67. Cited on page 907.

948 For some modern knot research, see P. Holdin, R.B. Kusner & A. Stasiak, Quantization of energy and writhe in self-repelling knots, *New Journal of Physics* **4**, pp. 20.1–20.11, 2002. No citations.

949 W.R. Taylor, A deeply knotted protein structure and how it might fold, *Nature* **406**, pp. 916–919, 2000. Cited on page 907.

950 A.C. Hirshfeld, Knots and physics: Old wine in new bottles, *American Journal of Physics* **66**, pp. 1060–1066, 1998. Cited on page 908.

951 T. Jacobson, Thermodynamics of spacetime: the Einstein equation of state, *Physical Review Letters* **75**, pp. 1260–1263, 1995, or http://www.arxiv.org/abs/gr-qc/9504004. Cited on page 910.

952 J. Frenkel & T. Kontorowa, Über die Theorie der plastischen Verformung, *Physikalische Zeitschrift der Sowietunion* **13**, pp. 1–10, 1938. F.C. Frank, On the equations of motion of crystal dislocations, *Proceedings of the Physical Society A* **62**, pp. 131–134, 1949, J. Eshelby, Uniformly moving dislocations, *Proceedings of the Physical Society A* **62**, pp. 307–314, 1949. See also G. Leibfried & H. Dietze, *Zeitschrift für Physik* **126**, p. 790, 1949. A general introduction can be found in A. Seeger & P. Schiller, Kinks in dislocations lines and their effects in internal friction in crystals, *Physical Acoustics* **3A**, W.P. Mason, ed., Academic Press, 1966. See also the textbooks by Frank R.N. Nabarro, *Theory of Crystal Dislocations*, Oxford University Press, 1967, or J.P. Hirth & J. Lothe, *Theory of Dislocations*, McGraw Hills Book Company, 1968. Cited on page 911.

953 J. Wisdom, Swimming in spacetime: motion by cyclic changes in body shape, *Science* **299**, pp. 1865–1869, 21st of March, 2003. Cited on page 912.

954 H.R. Pruppacher & J.D. Klett, *Microphysics of Clouds and Precipitation*, Reidel, 1978, pp. 316–319. Falling drops are flattened and look like a pill, due to the interplay between surface tension and air flow. Cited on page 913.

955 A manual for the ways to draw manifolds is the unique text by George Francis, *The Topological Picturebook*, Springer Verlag, 1987. It also contains a chapter on sphere eversion. Cited on page 915.

956 M.Yu. Konstantinov, Some problems of topology change description in the theory of space-time, http://www.arxiv.org/abs/gr-qc/9508050 preprint. No citations.

957 A. Unzicker, Teleparallel space-time with torsion yields geometrization of electrodynamcis with quantised charges, http://www.arxiv.org/abs/gr-qc/9612061 preprint. No citations.

Third Part

Motion Without Motion: What Are Space, Time and Particles?

Where through the combination of
quantum mechanics and general relativity,
the top of Motion Mountain is reached
and it is discovered
that vacuum is indistinguishable from matter,
that space, time and mass are easily confused,
that there is no difference between the very large and the very small,
and that a complete description of motion is possible.
(Well, wait a few more years for the last line.)

Chapter XI

General Relativity Versus Quantum Mechanics

The contradictions

> Man muß die Denkgewohnheiten durch Denknotwendigkeiten ersetzen.*
>
> Albert Einstein

The two stories told in the two parts of the path we have followed up to now, namely that on general relativity and the one on quantum field theory, are both beautiful and successful. Both are confirmed by experiments. We have reached a considerable height in our mountain ascent. The precision we achieved in the description of nature is impressive, and we are now able to describe all known examples of motion. So far we have encountered no exceptions.

However, the most important aspects of any type of motion, the masses of the particles involved and the strength of their coupling, are still unexplained. Furthermore, the origin of the number of particles in the universe, their initial conditions and the dimensionality of space-time remain hidden from us. Obviously, our adventure is not yet complete.

This last part of our hike will be the most demanding. In the ascent of any high mountain, the head gets dizzy because of the lack of oxygen. The finite energy at our disposal requires that we leave behind all unnecessary baggage and everything that slows us down. In order to determine what is unnecessary, we need to focus on what we want to achieve. Our biggest problem is all those concepts that are at the origin of the *contradictions* between general relativity and quantum theory. To pinpoint this useless baggage, we first list these contradictions.

In classical physics and in general relativity, the vacuum, or empty space-time, is a region with no mass, no energy and no momentum. If matter or gravitational fields are present, space-time is curved. The best way to measure the mass or energy content of
Ref. 958
space-time is to measure the average curvature of the universe. Cosmology tells us how
Page 424
we can do this; measurements yield an average energy density of the 'vacuum' of

$$E/V \approx 1\,\mathrm{nJ/m^3}\ . \qquad (640)$$

Ref. 959
However, quantum field theory tells a different story. Vacuum is a region with zero-point fluctuations. The energy content of vacuum is the sum of the zero-point energies of all the fields it contains. Indeed, the Casimir effect 'proves' the reality of these zero-point
Page 800
energies. Their energy density is given, within one order of magnitude, by

$$\frac{E}{V} = \frac{4\pi h}{c^3}\int_0^{\nu_{\max}} \nu^3 d\nu = \frac{\pi h}{c^3}\nu_{\max}^4\ . \qquad (641)$$

* 'One needs to replace habits of thought by necessities of thought.'

The approximation is valid for the case in which the cut-off frequency ν_{max} is much larger than the rest mass m of the particles corresponding to the field under consideration. Particle physicists argue that the cut-off energy has to be at least the energy of grand unification, about 10^{16} GeV= 1.6 MJ. That would give a vacuum energy density of

$$\frac{E}{V} \approx 10^{99}\,\mathrm{J/m^3}\,, \tag{642}$$

which is about 10^{108} times higher than the experimental limit deduced from spatial curvature using general relativity estimates. In other words, something is slightly wrong here.

General relativity and quantum theory contradict each other in other ways. Gravity is curved space-time. Extensive research has shown that quantum field theory, the description of electrodynamics and of nuclear forces, fails for situations with strongly curved space-times. In these cases the concept of 'particle' is not uniquely defined; quantum field theory cannot be extended to include gravity consistently and thus to include general relativity. Without the concept of the particle as a countable entity, the ability to perform perturbation calculations is also lost; and these are the only calculations possible in quantum field theory. In short, quantum theory only works because it assumes that gravity does not exist! Indeed, the gravitational constant does not appear in any consistent quantum field theory.

Ref. 960

On the other hand, general relativity neglects the commutation rules between physical quantities discovered in experiments on a microscopic scale. General relativity assumes that the position and the momentum of material objects can be given the meaning that they have in classical physics. It thus ignores Planck's constant $\hbar$ and only works by neglecting quantum theory.

Measurements also lead to problems. In general relativity, as in classical physics, it is assumed that infinite precision of measurement is possible, e.g. by using finer and finer ruler marks. In contrast, in quantum mechanics the precision of measurement is limited. The indeterminacy principle gives the limits that result from the mass M of the apparatus.

Page 608 Page 772

Time shows the contradictions most clearly. Relativity explains that time is what is read from clocks. Quantum theory says that precise clocks do not exist, especially if the coupling with gravitation is included. What does waiting 10 minutes mean, if the clock goes into a quantum mechanical superposition as a result of its coupling to space-time geometry?

In addition, quantum theory associates mass with an inverse length via the Compton wavelength; general relativity associates mass with length via the Schwarzschild radius.

Similarly, general relativity shows that space and time cannot be distinguished, whereas quantum theory says that matter does make a distinction. Quantum theory is a theory of – admittedly weirdly constructed – local observables. General relativity doesn't have any local observables, as Einstein's hole argument shows.

Page 457

Most dramatically, the contradiction is shown by the failure of general relativity to describe the pair creation of particles with spin 1/2, a typical and essential quantum process. John Wheeler and others have shown that, in such a case, the topology of space necessarily has to *change*; in general relativity, however, the topology of space is fixed. In short, quantum theory says that matter is made of fermions, while general relativity cannot in-

Ref. 961, Ref. 962 Ref. 963, Ref. 964

corporate fermions.

To sum up, general relativity and quantum theory clash. As long as an existing description of nature contains contradictions, it cannot lead to a unified description, to useful explanations, or even to a correct description. In order to proceed, let us take the shortest and fastest path: let us investigate the contradictions in more detail.

34. Does matter differ from vacuum?

Ref. 965 There is a simple way to state the origin of *all* contradictions between general relativity and quantum mechanics.* Both theories describe motion with objects made up of particles and with space-time made up of events. Let us see how these two concepts are defined.

A *particle* – and in general any object – is defined as a conserved entity to which a position can be ascribed and which can move. (The etymology of the term 'object' is connected to the latter fact.) In other words, a particle is a small entity with conserved mass, charge etc., which can vary its position with time.

Ref. 966 In every physics text *time* is defined with the help of moving objects, usually called 'clocks', or with the help of moving particles, such as those emitted by light sources. Similarly, the *length* is defined in terms of objects, either with an old-fashioned ruler or with the help of the motion of light, which in turn is motion of particles.

Modern physics has further sharpened the definitions of particle and space-time. Quantum mechanics assumes that space-time is given (it is included as a symmetry of the Hamiltonian), and studies the properties and the motion of particles, both for matter and for radiation. General relativity, and especially cosmology, takes the opposite approach: it assumes that the properties of matter and radiation are given, e.g. via their equations of state, and describes in detail the space-time that follows from them, in particular its curvature.

However, one fact remains unchanged throughout all these advances in physics: *the two concepts of particles and of space-time are each defined with the help of the other.* To avoid the contradiction between quantum mechanics and general relativity and to eliminate their incompleteness requires the elimination of this circular definition. As argued in the following, this necessitates a radical change in our description of nature, and in particular of the continuity of space-time.

For a long time, the contradictions between the two descriptions of nature were avoided by keeping them separate. One often hears the statement that quantum mechanics is valid at small dimensions and general relativity is valid at large dimensions, but this artificial separation is not justified; worse, it prevents the solution of the problem. The situation resembles the well-known drawing (Figure 364) by Maurits Escher (1898–1972) where two hands, each holding a pencil, seem to be drawing each other. If one hand is taken as a symbol of space-time and the other as a symbol of particles, with the act of drawing taken as a symbol of the act of defining, the picture gives a description of standard twentieth century physics.The apparent contradiction is solved by recognizing that

* The main results of this section are standard knowledge among specialists of unification; there are given here in simple arguments. For another way to derive the results, see the summary section on limit statements in nature, on page 985.

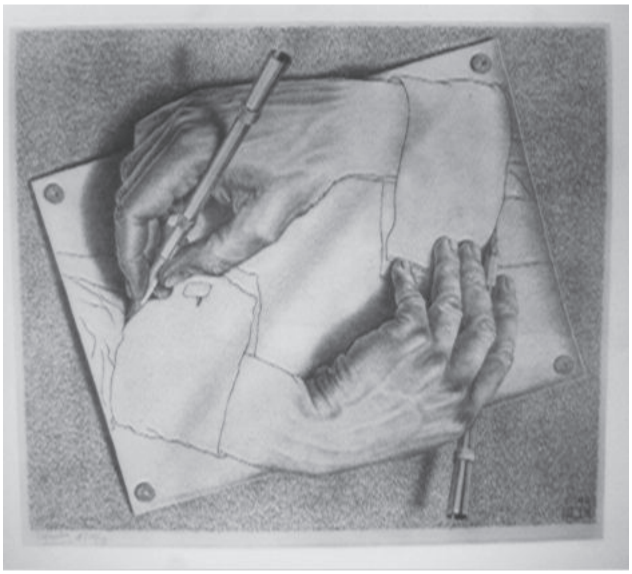


Figure 364 'Tekenen' by Maurits Escher, 1948 – a metaphor for the way in which 'particles' and 'space-time' are usually defined: each with the help of the other

the two concepts (the two hands) result from a third, hidden concept from which the other two originate. In the picture, this third entity is the hand of the painter.

Ref. 967, Ref. 968 Ref. 969, Ref. 970 In the case of space-time and matter, the search for the underlying common concept is presently making renewed progress. The required conceptual changes are so dramatic that they should be of interest to anybody who has an interest in physics. The most effective way to deduce the new concepts is to focus in detail on that domain where the contradiction between the two standard theories becomes most dramatic and where both theories are necessary at the same time. That domain is given by a well-known argument.

Planck scales

Both general relativity and quantum mechanics are successful theories for the description of nature. Each provides a criterion for determining when classical Galilean physics is no longer applicable. (In the following, we use the terms 'vacuum' and 'empty space-time' interchangeably.)

General relativity shows that it is necessary to take into account the curvature of space-time whenever we approach an object of mass m to within a distance of the order of the Schwarzschild radius r_S, given by

$$r_S = 2Gm/c^2 . \tag{643}$$

The gravitational constant G and the speed of light c act as conversion constants. Indeed, as the Schwarzschild radius of an object is approached, the difference between general relativity and the classical $1/r^2$ description of gravity becomes larger and larger. For example, the barely measurable gravitational deflection of light by the Sun is due to

Table 70 The size, Schwarzschild radius and Compton wavelength of some objects appearing in nature. The lengths between quotes make no physical sense, as explained in the text.

Object	Size: diam. d	Mass m	Schwarz-schild radius r_S	Ratio d/r_S	Compton wave-length λ_C	Ratio d/λ_C
galaxy	≈ 1 Zm	$\approx 5 \cdot 10^{40}$ kg	≈ 70 Tm	$\approx 10^{7}$	$\approx 10^{-83}$ m	$\approx 10^{104}$
neutron star	10 km	$2.8 \cdot 10^{30}$ kg	4.2 km	2.4	'$1.3 \cdot 10^{-73}$ m'	$8.0 \cdot 10^{76}$
Sun	1.4 Gm	$2.0 \cdot 10^{30}$ kg	3.0 km	$4.8 \cdot 10^{5}$	'$1.0 \cdot 10^{-73}$ m'	$8.0 \cdot 10^{81}$
Earth	13 Mm	$6.0 \cdot 10^{24}$ kg	8.9 mm	$1.4 \cdot 10^{9}$	'$5.8 \cdot 10^{-68}$ m'	$2.2 \cdot 10^{74}$
human	1.8 m	75 kg	0.11 ym	$1.6 \cdot 10^{25}$	'$4.7 \cdot 10^{-45}$ m'	$3.8 \cdot 10^{44}$
molecule	10 nm	0.57 zg	'$8.5 \cdot 10^{-52}$ m'	$1.2 \cdot 10^{43}$	$6.2 \cdot 10^{-19}$ m	$1.6 \cdot 10^{10}$
atom (^{12}C)	0.6 nm	20 yg	'$3.0 \cdot 10^{-53}$ m'	$2.0 \cdot 10^{43}$	$1.8 \cdot 10^{-17}$ m	$3.2 \cdot 10^{7}$
proton p	2 fm	1.7 yg	'$2.5 \cdot 10^{-54}$ m'	$8.0 \cdot 10^{38}$	$2.0 \cdot 10^{-16}$ m	9.6
pion π	2 fm	0.24 yg	'$3.6 \cdot 10^{-55}$ m'	$5.6 \cdot 10^{39}$	$1.5 \cdot 10^{-15}$ m	1.4
up-quark u	< 0.1 fm	0.6 yg	'$9.0 \cdot 10^{-55}$ m'	$< 1.1 \cdot 10^{38}$	$5.5 \cdot 10^{-16}$ m	< 0.18
electron e	< 4 am	$9.1 \cdot 10^{-31}$ kg	'$1.4 \cdot 10^{-57}$ m'	$3.0 \cdot 10^{39}$	$3.8 \cdot 10^{-13}$ m	$< 1.0 \cdot 10^{-5}$
neutrino ν_e	< 4 am	$< 3.0 \cdot 10^{-35}$ kg	'$< 4.5 \cdot 10^{-62}$ m'	n.a.	$> 1.1 \cdot 10^{-8}$ m	$< 3.4 \cdot 10^{-10}$

Ref. 961, Ref. 971 approaching it to within $2.4 \cdot 10^5$ times its Schwarzschild radius. Usually however, we are forced to stay away from objects at a distance that is an even larger multiple of the Schwarzschild radius, as shown in Table 70. For this reason, general relativity is unnecessary in everyday life. (An object smaller than its own Schwarzschild radius is called a *black hole*. According to general relativity, no signals from inside the Schwarzschild radius can reach the outside world; hence the name 'black hole'.) Ref. 972

Similarly, quantum mechanics shows that Galilean physics must be abandoned and quantum effects must be taken into account whenever an object is approached to within distances of the order of the (reduced) Compton wavelength λ_C, given by

$$\lambda_C = \frac{\hbar}{m\,c} \,. \tag{644}$$

In this case, Planck's constant h and the speed of light c act as conversion factors to transform the mass m into a length scale. Of course, this length only plays a role if the object itself is smaller than its own Compton wavelength. At these dimensions we get relativistic quantum effects, such as particle–antiparticle creation or annihilation. Table 70 shows that the approach distance is near or smaller than the Compton wavelength only in the microscopic world, so that such effects are not observed in everyday life. We do not therefore need quantum field theory to describe common observations.

The *combined* concepts of quantum field theory and general relativity are required in situations in which both conditions are satisfied simultaneously. The necessary approach distance for such situations is calculated by setting $r_S = 2\lambda_C$ (the factor 2 is introduced for simplicity). We find that this is the case when lengths or times are (of the order of)

$$\begin{aligned} l_{\mathrm{Pl}} = \sqrt{\hbar G/c^3} &= 1.6 \cdot 10^{-35}\ \mathrm{m}, \text{ the } \textit{Planck length,} \\ t_{\mathrm{Pl}} = \sqrt{\hbar G/c^5} &= 5.4 \cdot 10^{-44}\ \mathrm{s}\ , \text{ the } \textit{Planck time.} \end{aligned} \tag{645}$$

Whenever we approach objects at these scales, both general relativity and quantum mechanics play a role; at these scales effects of *quantum gravity* appear. Because the values of the Planck dimensions are extremely small, this level of sophistication is unnecessary in everyday life, in astronomy and even in particle physics.

However, to answer the questions posted at the beginning of the book – why do we live in three dimensions and why is the proton 1836.15 times heavier than the electron? – we require a precise and complete description of nature. The contradictions between quantum mechanics and general relativity appear to make the search for answers impossible. However, while the unified theory describing quantum gravity is not yet complete, we can already get a few glimpses at its implications from its present stage of development.

Note that the Planck scales specify one of only two domains of nature where quantum Challenge 1370 n mechanics and general relativity apply at the same time. (What is the other?) As Planck scales are the easier of the two to study, they provide the best starting point for the fol- Ref. 973 lowing discussion. When Max Planck discovered the existence of Planck scales or Planck units, he was interested in them mainly as *natural units of measurement*, and that is what he called them. However, their importance in nature is much more widespread, as we will shall see in the new section. We will discover that they determine what is commonly called *quantum geometry*.

Farewell to instants of time

> Time is composed of time atoms ... which in fact are indivisible.
>
> Moses Maimonides, 12th century.

The appearance of the quantum of action in the description of motion leads to quantum limits to all measurements. These limits have important consequences at Planck dimensions. Measurement limits appear most clearly when we investigate the properties of clocks and meter rules. Is it possible to construct a clock that is able to measure time Ref. 974, Ref. 975 intervals shorter than the Planck time? Surprisingly, the answer is no, even though the time–energy indeterminacy relation $\Delta E \Delta t \geqslant \hbar$ seems to indicate that by making ΔE arbitrary large, we can make Δt arbitrary small.

Every clock is a device with some moving parts. Parts can be mechanical wheels, particles of matter in motion, changing electrodynamic fields, i.e. photons, or decaying Ref. 976, Ref. 977 radioactive particles. For each moving component of a clock, such as the two hands, the Ref. 978 indeterminacy principle applies. As discussed most clearly by Michael Raymer, the indeterminacy relation for two non-commuting variables describes two different, but related situations: it makes a statement about standard *deviations* of *separate* measurements on *many* identical systems; and it describes the measurement *precision* for a *joint* measurement on a *single* system. Throughout this article, only the second situation is considered.

For any clock to work, we need to know both the time and the energy of each hand. Otherwise it would not be a recording device. Put more generally, a clock must be a classical system. We need the combined knowledge of the non-commuting variables for each moving component of the clock. Let us focus on the component with the largest time indeterminacy Δt. It is evident that the smallest time interval δt that can be measured by a clock is always larger than the quantum limit, i.e. larger than the time indeterminacy Δt for the most 'uncertain' component. Thus we have

$$\delta t \geqslant \Delta t \geqslant \frac{\hbar}{\Delta E} , \tag{646}$$

where ΔE is the energy indeterminacy of the moving component, and this energy indeterminacy ΔE must be smaller than the total energy $E = mc^2$ of the component itself.* Furthermore, a clock provides information and thus signals have to be able to leave it. To make this possible, the clock must not be a black hole and its mass m must therefore be smaller than the Schwarzschild mass for its size, i.e. $m \leqslant c^2 l/G$, where l is the size of the clock (neglecting factors of order unity). Finally, for a sensible measurement of the time interval δt, the size l of the clock must be smaller than $c\,\delta t$ itself, because otherwise different parts of the clock could not work together to produce the same time display.** If we combine all these conditions, we get

$$\delta t \geqslant \frac{\hbar G}{c^5 \delta t} \tag{647}$$

or

$$\delta t \geqslant \sqrt{\frac{\hbar G}{c^5}} = t_{\mathrm{Pl}} . \tag{648}$$

In summary, from three simple properties of any clock, namely that there is only a single clock, that we can read its dial and that it gives sensible read-outs, we get the general conclusion that *clocks cannot measure time intervals shorter than the Planck time*. Note that this argument is independent of the nature of the clock mechanism. Whether the clock is powered by gravitational, electrical, plain mechanical or even nuclear means, the limit still applies.***

Ref. 982 The same result can also be found in other ways. For example, any clock small enough to measure small time intervals necessarily has a certain energy indeterminacy due to

* Physically, this condition means being sure that there is only *one* clock; the case $\Delta E > E$ would mean that it is impossible to distinguish between a single clock and a clock–anticlock pair created from the vacuum, or a component plus two such pairs, etc.

** It is amusing to explore how a clock larger than $c\,\delta t$ would stop working, as a result of the loss of rigidity in its components. Challenge 1371 n

*** Note that gravitation is essential here. The present argument differs from the well-known study on the limitations of clocks due to their mass and their measuring time which was published by Salecker and Wigner and summarized in pedagogical form by Zimmerman. Here, both quantum mechanics and gravity are included, and therefore a different, lower and much more fundamental limit is found. Note also that the discovery of black hole radiation does not change the argument; black hole radiation notwithstanding, measurement devices cannot exist inside black holes.

Ref. 980 Ref. 981

the indeterminacy relation. At the same time, on the basis of general relativity, any energy density induces a deformation of space-time and signals from the deformed region arrive
Page 376
with a certain delay due to that deformation. The energy indeterminacy of the source leads to an indeterminacy in the deformation and thus in the delay. The expression from
Ref. 971
general relativity for the deformation of the time part of the line element due to a mass m is $\delta t = mG/lc^3$. From the mass–energy relation, an energy spread ΔE produces an indeterminacy Δt in the delay

$$\Delta t = \frac{\Delta E\, G}{l\, c^5} . \tag{649}$$

This indeterminacy determines the precision of the clock. Furthermore, the energy indeterminacy of the clock is fixed by the indeterminacy relation for time and energy $\Delta E \geqslant \hbar/\Delta t$, in turn fixed by the precision of the clock. Combining all this, we again find the relation $\delta t \geqslant t_{\mathrm{Pl}}$ for the minimum measurable time. We are forced to conclude that *in nature there is a minimum time interval.* In other words, *at Planck scales the term 'instant of time' has no theoretical or experimental basis.* It therefore makes no sense to use the term.

Farewell to points in space

In a similar way, we can deduce that it is impossible to make a meter rule or any other
Ref. 983
length measuring device that is able to measure lengths shorter than the Planck length. Obviously, we can already deduce this from $l_{\mathrm{Pl}} = c\, t_{\mathrm{Pl}}$, but a separate proof is also possible.

The straightforward way to measure the distance between two points is to put an object at rest at each position. In other words, joint measurements of position and momentum are necessary for every length measurement. Now, the minimal length δl that can be measured must be larger than the position indeterminacy of the two objects. From the indeterminacy principle it is known that each object's position cannot be determined with a precision Δl better than that given by the indeterminacy relation $\Delta l\, \Delta p = \hbar$, where Δp is the momentum indeterminacy. The requirement that there is only one object at each end, i.e. avoiding pair production from the vacuum, means that $\Delta p < mc$; together, these requirements give

$$\delta l \geqslant \Delta l \geqslant \frac{\hbar}{mc} . \tag{650}$$

Furthermore, the measurement cannot be performed if signals cannot leave the objects; thus they may not be black holes. Therefore their masses must be small enough for their Schwarzschild radius $r_{\mathrm{S}} = 2Gm/c^2$ to be smaller than the distance δl separating them. Again omitting the factor of 2, we get

$$\delta l \geqslant \sqrt{\frac{\hbar G}{c^3}} = l_{\mathrm{Pl}} . \tag{651}$$

Another way to deduce this limit reverses the roles of general relativity and quantum theory. To measure the distance between two objects, we have to localize the first object with respect to the other within a certain interval Δx. The corresponding energy indeterminacy obeys $\Delta E = c(c^2m^2 + (\Delta p)^2)^{1/2} \geqslant c\hbar/\Delta x$. However, general relativity shows that a small volume filled with energy changes the curvature of space-time, and thus changes the metric of the surrounding space. For the resulting distance change Δl, compared to empty space, we find the expression $\Delta l \approx G\Delta E/c^4$. In short, if we localize the first particle in space with a precision Δx, the distance to a second particle is known only with precision Δl. The minimum length δl that can be measured is obviously larger than either of these quantities; inserting the expression for ΔE, we find again that the minimum measurable length δl is given by the Planck length.

Ref. 961, Ref. 971

Ref. 983, Ref. 984, Ref. 985

Ref. 986, Ref. 987, Ref. 988

We note that, as the Planck length is the shortest possible length, it follows that there can be no observations of quantum mechanical effects for situations in which the corresponding de Broglie or Compton wavelength is smaller than the Planck length. In proton–proton collisions we observe both pair production and interference effects. In contrast, the Planck limit implies that in everyday, macroscopic situations, such as car–car collisions, we cannot observe embryo–antiembryo pair production and quantum interference effects.

In summary, from two simple properties common to all length measuring devices, namely that they can be counted and that they can be read out, we arrive at the conclusion that *lengths smaller than the Planck length cannot be found in measurements*. Whatever method is used, be it a meter rule or time-of-flight measurement, we cannot overcome this fundamental limit. It follows that *the concept of a 'point in space' has no experimental basis*. In the same way, the term 'event', being a combination of a 'point in space' and an 'instant of time', also loses its meaning for the description of nature.

A simple way to deduce the minimum length using the limit statements which structure this ascent is the following. General relativity is based on a maximum force in nature, or alternatively, on a minimum mass change per time; its value is given by $\mathrm{d}m/\mathrm{d}t = c^3/4G$. Quantum theory is based on a minimum action in nature, given by $L = \hbar/2$. Since a distance d can be expressed like

Page 319

$$d^2 = \frac{L}{\mathrm{d}m/\mathrm{d}t} , \tag{652}$$

one sees directly that a minimum action and a maximum mass rate imply a minimum distance. In other words, quantum theory and general relativity, when put together, imply a minimum distance.

Ref. 989

These results are often expressed by the so-called *generalized indeterminacy principle*

$$\Delta p \Delta x \geqslant \hbar/2 + f \frac{G}{c^3} (\Delta p)^2 \tag{653}$$

or

$$\Delta p \Delta x \geqslant \hbar/2 + f \frac{l_{\mathrm{Pl}}^2}{\hbar} (\Delta p)^2 , \tag{654}$$

where f is a numerical factor of order unity. A similar expression holds for the time–energy indeterminacy relation. The first term on the right hand side is the usual quantum mechanical indeterminacy. The second term, negligible for everyday life energies, plays a role only near Planck energies and is due to the changes in space-time induced by gravity

Challenge 1372 e at these high energies. You should be able to show that the generalized principle (653) automatically implies that Δx can never be smaller than $f^{1/2} l_{\text{Pl}}$.

The generalized indeterminacy principle is derived in exactly the same way in which Heisenberg derived the original indeterminacy principle $\Delta p \Delta x \geqslant \hbar/2$, namely by studying the deflection of light by an object under a microscope. A careful re-evaluation of the process, this time including gravity, yields equation (653). For this reason, *all* approaches that try to unify quantum mechanics and gravity must yield this relation; indeed, it appears in the theory of canonical quantum gravity, in superstring theory and in the quantum group approach. Ref. 989 Ref. 990 Ref. 991 Ref. 992, Ref. 993, Ref. 994

We remember that quantum mechanics starts when we realize that the classical concept of action makes no sense below the value of $\hbar/2$; similarly, unified theories start when we realize that the classical concepts of time and length make no sense below Planck values. However, the usual description of space-time does contain such small values; the usual description involves the existence of intervals smaller than the smallest measurable one. *Therefore, the continuum description of space-time has to be abandoned in favour of a more appropriate description.*

Ref. 995 The new indeterminacy relation appearing at Planck scales shows that continuity cannot be a good description of space-time. Inserting $c\Delta p \geqslant \Delta E \geqslant \hbar/\Delta t$ into equation (653), we get

$$\Delta x \Delta t \geqslant \hbar G/c^4 = t_{\text{Pl}} l_{\text{Pl}} , \tag{655}$$

which of course has no counterpart in standard quantum mechanics. It shows that space-time events do not exist. A final way to convince oneself that points have no meaning is that a point is an entity with vanishing volume; however, the minimum volume possible in nature is the Planck volume $V_{\text{Pl}} = l_{\text{Pl}}^3$.

While space-time points are idealizations of events, this idealization is incorrect. The use of the concept of 'point' is similar to the use of the concept of 'aether' a century ago: it is impossible to detect and it is only useful for describing observations until a way to describe nature without it has been found. Like 'aether´, also 'point´ leads reason astray.

In other words, the Planck units do not only provide natural units, they also provide — within a factor of order one — the *limit* values of space and time intervals.

Farewell to the space-time manifold

The consequences of the Planck limits for measurements of time and space can be taken much further. It is commonplace to say that given any two points in space or any two instants of time, there is always a third in between. Physicists sloppily call this property continuity, while mathematicians call it denseness. However, at Planck dimensions this property cannot exist, since intervals smaller than the Planck time can never be found. Thus points and instants are not dense, and *between two points there is not always a third.* This means that *space and time are not continuous.* Of course, at large scales they are – approximately – continuous, in the same way that a piece of rubber or a liquid seems continuous at everyday dimensions, even though it is not at a small scale.

All paradoxes resulting from the infinite divisibility of space and time, such as Zeno's argument on the impossibility to distinguish motion from rest, or the Banach–Tarski paradox, are now avoided. We can dismiss the paradoxes straight away because of their incorrect premises concerning the nature of space and time.

But let us go on. Special relativity, quantum mechanics and general relativity all rely on the idea that time can be defined for all points of a given reference frame. However, two clocks a distance l apart cannot be synchronized with arbitrary precision. Since the distance between two clocks cannot be measured with an error smaller than the Planck length l_{Pl}, and transmission of signals is necessary for synchronization, it is not possible to synchronize two clocks with a better precision than the time $l_{Pl}/c = t_{Pl}$, the Planck time. Because it is impossible to synchronize clocks precisely, a single time coordinate for a whole reference frame is only an approximation, and this idea cannot be maintained in a precise description of nature.

Moreover, since the time difference between events can only be measured within a Planck time, for two events distant in time by this order of magnitude, it is not possible to say with complete certainty which of the two *precedes* the other! This is an important result. If events cannot be ordered, the concept of time, which was introduced into physics to describe sequences, cannot be defined at all at Planck scales. In other words, after dropping the idea of a common time coordinate for a complete frame of reference, we are forced to drop the idea of time at a single 'point' as well. Therefore, the concept of 'proper time' loses its meaning at Planck scales.

Ref. 996

It is straightforward to use the same arguments to show that length measurements do not allow us to speak of continuous space, but only of *approximately* continuous space. As a result of the lack of measurement precision at Planck scales, the concepts of spatial order, translation invariance, isotropy of vacuum and global coordinate systems have no experimental basis.

But there is more to come. The very existence of a minimum length contradicts special relativity theory, in which it is shown that lengths undergo Lorentz contraction when the frame of reference is changed. A minimum length thus cannot exist in special relativity. But we just deduced that there must be such a minimum distance in nature. There is only one conclusion: special relativity cannot be correct at smallest distances. Thus, *space-time is neither Lorentz invariant nor diffeomorphism invariant nor dilatation invariant at Planck dimensions.* All symmetries that are at the basis of special and general relativity are thus only approximately valid at Planck scales.

As a result of the imprecision of measurement, most familiar concepts used to describe spatial relations become useless. For example, the concept of *metric* loses its usefulness at Planck scales. Since distances cannot be measured with precision, the metric cannot be determined. We deduce that it is impossible to say precisely whether space is flat or curved. In other words, *the impossibility of measuring lengths exactly is equivalent to fluctuations of the curvature, and thus equivalent to fluctuations of gravity*.

Ref. 983, Ref. 997

In addition, even the number of spatial dimensions makes no sense at Planck scales. Let us remind ourselves how to determine this number experimentally. One possible way is to determine how many points we can choose in space such that all the distances between them are equal. If we can find at most n such points, the space has $n-1$ dimensions. We can see that if reliable length measurement at Planck scale is not possible, there is no way to determine reliably the number of dimensions of a space with this method.

Another way to check for three spatial dimensions is to make a knot in a shoe string and glue the ends together: since it stays knotted we know that space has three dimensions, because there is a mathematical theorem that in spaces with greater or fewer than three dimensions, knots do not exist. Again, at Planck dimensions the errors in measurement do not allow to say whether a string is knotted or not, because measurement limits at crossings make it impossible to say which strand lies above the other; in short, at Planck scales we cannot check whether space has three dimensions or not.

There are many other methods for determining the dimensionality of space.* All these methods start from the definition of the concept of dimensionality, which is based on a precise definition of the concept of neighbourhood. However, at Planck scales, as just mentioned, length measurements do not allow us to say whether a given point is inside or outside a given volume. In short, whatever method we use, the lack of reliable length measurements means that *at Planck scales, the dimensionality of physical space is not defined.* It should therefore not come as a surprise that when we *approach* these scales, we may get a scale-dependent answer for the number of dimensions, that may be different from three.

The reason for the problems with space-time become most evident when we remember Euclid's well-known definition: 'A point is that which has no part.' As Euclid clearly understood, a physical point, and here the stress is on *physical*, cannot be defined *without* some measurement method. A physical point is an idealization of position, and as such includes measurement right from the start. In mathematics, however, Euclid's definition is rejected; mathematical points do not need a metric for their definition. Mathematical points are elements of a set, usually called a space. In mathematics, a measurable or metric space is a set of points equipped *afterwards* with a measure or a metric. Mathematical points do not need a metric for their definition; they are basic entities. In contrast to the mathematical situation, the case of physical space-time, the concepts of measure and of metric are *more fundamental* than that of a point. The difficulties distinguishing physical and mathematical space and points arise from the failure to distinguish a mathematical metric from a physical length measurement.**

Ref. 998

Perhaps the most beautiful way to make this point is the Banach–Tarski theorem, which clearly shows the limits of the concept of volume. The theorem states that a sphere made up of *mathematical points* can be cut into five pieces in such a way that the pieces

Ref. 999

* For example, we can determine the dimension using only the topological properties of space. If we draw a so-called *covering* of a topological space with open sets, there are always points that are elements of several sets of the covering. Let us call p the maximal number of sets of which a point can be an element in a given covering. This number can be determined for all possible coverings. The minimum value of p, minus one, gives the dimension of the space.

In fact, if physical space is not a manifold, the various methods may give different answers for the dimensionality. Indeed, for linear spaces without norm, a unique number of dimensions cannot be defined. The value then depends on the specific definition used and is called e.g. fractal dimension, Lyapunov dimension, etc.

** Where does the incorrect idea of continuous space-time have its roots? In everyday life, as well as in physics, space-time is introduced to describe observations. Space-time is a book-keeping device. Its properties are extracted from the properties of observables. Since observables can be added and multiplied, we extrapolate that they can take continuous values. This extrapolation implies that length and time intervals can take continuous values, and, in particular, arbitrary small values. From this result we get the possibility of defining points and sets of points. A special field of mathematics, topology, shows how to start from a set of points and construct, with the help of neighbourhood relations and separation properties, first a *topological space*. Then, with the help of a metric, a *metric space* can be built. With the appropriate compactness and connectedness relations, a *manifold*, characterized by its dimension, metric and topology, can be constructed.

can be put together to form two spheres, each of the *same volume* as the original one. However, the necessary cuts are 'infinitely' curved and detailed: they are wildly disconnected. For physical matter such as gold, unfortunately – or fortunately – the existence of a minimum length, namely the atomic distance, makes it impossible to perform such a cut. For vacuum, the puzzle reappears: for example, the energy of zero-point fluctuations is given by the density times the volume; following the Banach–Tarski theorem, the zero point energy content of a single sphere should be equal to the zero point energy of two similar spheres each of the same volume as the original one. The paradox is solved by the Planck length, because it also provides a fundamental length scale for vacuum, thus making infinitely complex cuts impossible. Therefore, *the concept of volume is only well defined at Planck scales if a minimum length is introduced*.

To sum up, *physical space-time cannot be a set of mathematical points.* But the surprises are not finished. At Planck dimensions, since both temporal and spatial order break down, there is no way to say if the distance between two space-time regions that are close enough together is space-like or time-like. Measurement limits make it impossible to distinguish the two cases. *At Planck scales, time and space cannot be distinguished from each other*.

In addition, it is impossible to state that the topology of space-time is fixed, as general relativity implies. The topology changes – mentioned above – required for particle reactions do become possible. In this way another of the contradictions between general relativity and quantum theory is resolved.

In summary, space-time at Planck scales is not continuous, not ordered, not endowed with a metric, not four-dimensional and not made up of points. If we compare this with the definition of the term manifold,* not one of its defining properties is fulfilled. We arrive at the conclusion that *the concept of a space-time manifold has no backing at Planck scales*. This is a strong result. *Even though both general relativity and quantum mechanics use continuous space-time, the combination of both theories does not.*

> There is nothing in the world but matter in motion, and matter in motion cannot move otherwise than in space and time.
>
> Lenin

Farewell to observables and measurements

To complete this review of the situation, if space and time are not continuous, no quantities defined as derivatives with respect to space or time are precisely defined. Velocity, acceleration, momentum, energy, etc., are only well-defined under the assumption of continuous space and time. That important tool, the evolution equation, is based on derivatives and thus can no longer be used. Therefore the Schrödinger or the Dirac equation lose their basis. Concepts such as 'derivative', 'divergence-free', 'source free' etc., lose their meaning at Planck scales.

In fact, all physical observables are defined using length and time measurements. A list of physical units shows that each is a product of powers of length, time (and mass) units. (Even though in the SI system electrical quantities have a separate base quantity,

* A manifold is what *locally* looks like an Euclidean space. The exact definition can be found in Appendix D.

the ampere, the argument still holds; the ampere is itself defined in terms of a force, which is measured using the three base units length, time and mass.) Since time and length are not continuous, observables themselves are not defined, because their value is not fixed. This means that *at Planck scales, observables cannot be described by real numbers.*

In addition, if time and space are not continuous, the usual expression for an observable field A, namely $A(t, x)$, does not make sense: we have to find a more appropriate description. *Physical fields cannot exist at Planck scales.*

The consequences for quantum mechanics are severe. It makes no sense to define multiplication of observables by real numbers, thus by a continuous range of values, but only by a discrete set of numbers. Among other implications, this means that observables do not form a linear algebra. We recognize that, because of measurement errors, we cannot prove that observables do form such an algebra. This means that *observables are not described by operators at Planck scales.* And, because quantum mechanics is based on the superposition principle, without it, everything comes crumbling down. In particular, the most important observables are the gauge potentials. Since they do not now form an algebra, *gauge symmetry is not valid at Planck scales*. Even innocuous looking expressions such as $[x_{\mathrm{i}}, x_{\mathrm{j}}] = 0$ for $x_{\mathrm{i}} \neq x_{\mathrm{j}}$, which are at the root of quantum field theory, become meaningless at Planck scales. Since at those scales also the superposition principle cannot be backed up by experiment, even the famous Wheeler–DeWitt equation, often assumed to describe quantum gravity, cannot be valid.

Similarly, permutation symmetry is based on the premise that we can distinguish two points by their coordinates, and then exchange particles between those two locations. As we have just seen, this is not possible if the distance between the two particles is small; we conclude that *permutation symmetry has no experimental basis at Planck scales*.

Even discrete symmetries, like charge conjugation, space inversion and time reversal cannot be correct in this domain, because there is no way to verify them exactly by measurement. *CPT symmetry is not valid at Planck scales.*

Finally we note that all types of scaling relations do not work at smallest scales. *As a result, renormalization symmetry is also destroyed at Planck scales.*

All these results are consistent: if there are no symmetries at Planck scales, there are also no observables, since physical observables are representations of symmetry groups. In fact, the limits on time and length measurements imply that *the concept of measurement has no significance at Planck scales*.

Can space-time be a lattice? – A glimpse of quantum geometry

Ref. 1000 Discretization of space-time has been studied already since 1940s. Recently, the idea that
Ref. 1001 space-time could be described as a lattice has also been explored most notably by David
Ref. 1002 Finkelstein and by Gerard 't Hooft. The idea of space-time as a lattice is based on the idea
Ref. 1004 that, if a minimum distance exists, then all distances are a multiple of this minimum. It is generally agreed that, in order to get an isotropic and homogeneous situation for large,
Ref. 1003 everyday scales, the structure of space-time cannot be periodic, but must be random. In addition, any fixed structure of space-time violates the result that there are no lengths smaller than the Planck length: as a result of the Lorentz contraction, any moving observer would find lattice distances smaller than the Planck value. Worse still, the lattice idea conflicts with general relativity, in particular with the diffeomorphism invariance of

vacuum. Finally, where would a particle be *during* the jump from one lattice point to the next? Thus, in summary, *space-time cannot be a lattice.* A minimum distance does exist in nature; however, the hope that all other distances are simple multiples of the smallest distance is not correct. We will discover more evidence for this later on.

If space-time is not a set of points or events, it must be something else. Three hints already appear at this stage. The first step necessary to improve the description of motion is the recognition that abandoning 'points' means abandoning the *local* description of nature. Both quantum mechanics and general relativity assume that the phrase 'observable at a point' has a precise meaning. Because it is impossible to describe space as a manifold, this expression is no longer useful. The unification of general relativity and quantum physics forces the adoption of a *non-local* description of nature at Planck scales.

The existence of a minimum length implies that there is no way to physically distinguish locations that are even closer together. We are tempted to conclude therefore that *no* pair of locations can be distinguished, even if they are one metre apart, since on any path joining two points, no two locations that are close together can be distinguished. This situation is similar to the question about the size of a cloud or of an atom. If we measure water density or electron density, we find non-vanishing values at any distance from the centre of the cloud or the atom; however, an effective size can still be defined, because it is very unlikely that the effects of the presence of a cloud or of an atom can be seen at distances much larger than this effective size. Similarly, we can guess that two points in space-time at a macroscopic distance from each other can be distinguished because the probability that they will be confused drops rapidly with increasing distance. In short, we are thus led to a *probabilistic* description of space-time. Space-time becomes a macroscopic observable, a *statistical* or *thermodynamic limit* of some microscopic entities.

We note that a fluctuating structure for space-time would also avoid the problems of fixed structures with Lorentz invariance. This property is of course compatible with a statistical description. In summary, the experimental observations of special relativity, i.e. Lorentz invariance, isotropy and homogeneity, together with that of a minimum distance, point towards a *fluctuating* description of space-time. Research efforts in quantum
Ref. 1005
gravity, superstring theory and quantum groups have confirmed independently of each other that a probabilistic and non-local description of space-time at Planck dimensions, resolves the contradictions between general relativity and quantum theory. This is our first result on quantum geometry. To clarify the issue, we have to turn to the concept of the particle.

Farewell to particles

In every example of motion, some object is involved. One of the important discoveries of the natural sciences was that all objects are composed of small constituents, called *elementary particles*. Quantum theory shows that all composite, non-elementary objects have a finite, non-vanishing size. This property allows us to determine whether a particle is elementary or not. If it behaves like a point particle, it is elementary. At present,only the leptons (electron, muon, tau and the neutrinos), the quarks and the radiation quanta of the electromagnetic, weak and strong nuclear interactions (the photon, the W and Z
Page 1078
bosons, the gluons) have been found to be elementary. A few more elementary particles

are predicted by various refinements of the standard model. Protons, atoms, molecules, cheese, people, galaxies etc., are all composite, as shown in Table 70. Elementary particles are characterized by their vanishing size, their spin and their mass.

Page 922

Even though the definition of 'elementary particle' as point particle is all we need in the following argument, it is not complete, because it seems to leave open the possibility that future experiments could show that electrons or quarks are not elementary. This is not so! In fact, any particle smaller than its own Compton wavelength is elementary. If it were composite, there would be a lighter component inside it and this lighter particle would have a larger Compton wavelength than the composite particle. This is impossible, since the size of a composite particle must be larger than the Compton wavelength of its components. (The possibility that all components are heavier than the composite, which would avoid this argument, does not lead to satisfying physical properties; for example, it leads to intrinsically unstable components.)

The *size* of an object, such as those given in Table 70, is defined as the length at which differences from point-like behaviour are observed. This is the way in which, using alpha particle scattering, the radius of the atomic nucleus was determined for the first time in Rutherford's experiment. In other words, the size d of an object is determined by measuring how it scatters a beam of probe particles. In daily life as well, when we look at objects, we make use of scattered photons. In general, in order to make use of scattering, the effective wavelength $\lambda = \hbar/mv$ of the probe must be smaller than the object size d to be determined. We thus need $d > \lambda = \hbar/(mv) \geqslant \hbar/(mc)$. In addition, in order to make a scattering experiment possible, the object must not be a black hole, since, if it were, it would simply swallow the approaching particle. This means that its mass m must be smaller than that of a black hole of the same size; in other words, from equation (643) we must have $m < dc^2/G$. Combining this with the previous condition we get

$$d > \sqrt{\frac{\hbar G}{c^3}} = l_{\mathrm{Pl}} . \tag{656}$$

In other words, there is no way to observe that an object is smaller than the Planck length. *There is thus no way in principle to deduce from observations that a particle is point-like.* In fact, it makes no sense to use the term 'point particle' at all! Of course, there is a relation between the existence of a minimum length for empty space and a minimum length for objects. If the term 'point of space' is meaningless, then the term 'point particle' is also meaningless. As in the case of time, the lower limit on length results from the combination of quantum mechanics and general relativity.*

The size d of any elementary particle must by definition be smaller than its own Compton wavelength $\hbar/(mc)$. Moreover, the size of a particle is always larger than the Planck length: $d > l_{\mathrm{Pl}}$. Combining these two requirements and eliminating the size d we get the condition for the mass m of any elementary particle, namely

$$m < \frac{\hbar}{c\, l_{\mathrm{Pl}}} = \sqrt{\frac{\hbar c}{G}} = m_{\mathrm{Pl}} = 2.2 \cdot 10^{-8}\,\mathrm{kg} = 1.2 \cdot 10^{19}\,\mathrm{GeV}/\mathrm{c}^2 . \tag{657}$$

* Obviously, the minimum size of a particle has nothing to do with the impossibility, in quantum theory, of localizing a particle to within less than its Compton wavelength.

The limit $m_{\rm Pl}$, the so-called *Planck mass*, corresponds roughly to the mass of a human embryo that is ten days old, or equivalently, to that of a small flea. In short, *the mass of any elementary particle must be smaller than the Planck mass.* This fact was already noted as 'well-known' by Andrei Sakharov* in 1968; he explains that these hypothetical particles are sometimes called 'maximons'. And indeed, the known elementary particles all have masses well below the Planck mass. (In fact, the question why their masses are so incredibly much smaller than the Planck mass is one of the most important questions of high-energy physics. We will come back to it.)

Ref. 1006

There are many other ways to arrive at the mass limit for particles. For example, in order to measure mass by scattering – and that is the only way for very small objects – the Compton wavelength of the scatterer must be larger than the Schwarzschild radius; otherwise the probe will be swallowed. Inserting the definition of the two quantities and neglecting the factor 2, we get again the limit $m < m_{\rm Pl}$. (In fact it is a general property of descriptions of nature that a minimum space-time interval leads to an upper limit for elementary particle masses.) The importance of the Planck mass will become clear shortly.

Ref. 1007

Figure 365 Andrei Sakharov

Another property connected with the size of a particle is its electric dipole moment. It describes the deviation of its charge distribution from spherical. Some predictions from the standard model of elementary particles give as *upper* limit for the electron dipole moment d_e a value of

Ref. 1008

$$\frac{|d_e|}{e} < 10^{-39}\,\mathrm{m}\,, \tag{658}$$

where e is the charge of the electron. This value is ten thousand times smaller than the Planck length $l_{\rm Pl}\,e$. Since the Planck length is the smallest possible length, we seem to have a potential contradiction here. However, a more recent prediction from the standard model is more careful and only states

Ref. 1009

$$\frac{|d_e|}{e} < 3\cdot 10^{-21}\,\mathrm{m}\,, \tag{659}$$

which is not in contradiction with a minimal length in nature. The issue is still not settled. We will see below that the experimental limit is expected to allow to test these predictions in the foreseeable future.

Planck scales have other strange consequences. In quantum field theory, the difference between a virtual particle and a real particle is that a real particle is 'on shell', obeying $E^2 = m^2c^4 + p^2c^2$, whereas a virtual particle is 'off shell', obeying $E^2 \neq m^2c^4 + p^2c^2$. Because of the fundamental limits of measurement precision, *at Planck scales we cannot determine whether a particle is real or virtual.*

* Andrei Dmitrievich Sakharov, famous Soviet nuclear physicist (1921–1989). One of the keenest thinkers in physics, Sakharov, among others, invented the Tokamak, directed the construction of nuclear bombs, and explained the matter-antimatter asymmetry of nature. Like many others, he later campaigned against nuclear weapons, a cause for which he was put into jail and exile, together with his wife, Yelena Bonner. He received the Nobel Peace Prize in 1975.

However, that is not all. Antimatter can be described as matter moving backwards in time. Since the difference between backwards and forwards cannot be determined at Planck scales, *matter and antimatter cannot be distinguished at Planck scales.*

Particles are also characterized by their spin. Spin describes two properties of a particle: its behaviour under rotations (and if the particle is charged, its behaviour in magnetic fields) and its behaviour under particle exchange. The wave function of particles with spin 1 remains invariant under a rotation of 2π, whereas that of particles with spin 1/2 changes sign. Similarly, the combined wave function of two particles with spin 1 does not change sign under exchange of particles, whereas for two particles with spin 1/2 it does.

We see directly that both transformations are impossible to study at Planck scales. Given the limit on position measurements, the position of a rotation axis cannot be well defined, and rotations become impossible to distinguish from translations. Similarly, position imprecision makes impossible the determination of precise separate positions for exchange experiments. In short, *spin cannot be defined at Planck scales, and fermions cannot be distinguished from bosons, or, phrased differently, matter cannot be distinguished from radiation at Planck scales.* We can thus easily see that supersymmetry, a unifying symmetry between bosons and fermions, somehow becomes natural at Planck dimensions.

But let us now move to the main property of elementary particles.

Farewell to mass

The Planck mass divided by the Planck volume, i.e. the Planck density, is given by

$$\rho_{\mathrm{Pl}} = \frac{c^5}{G^2 \hbar} = 5.2 \cdot 10^{96}\ \mathrm{kg/m^3} \tag{660}$$

and is a useful concept in the following. If we want to measure the (gravitational) mass M enclosed in a sphere of size R and thus (roughly) of volume R^3, one way to do this is to put a test particle in orbit around it at that same distance R. Universal gravitation then gives for the mass M the expression $M = Rv^2/G$, where v is the speed of the orbiting test particle. From $v < c$, we thus deduce that $M < c^2 R/G$; since the minimum value for R is the Planck distance, we get (neglecting again factors of order unity) a limit for the mass density ρ, namely

Challenge 1373 e

$$\rho < \rho_{\mathrm{Pl}} . \tag{661}$$

In other words, *the Planck density is the maximum possible value for mass density.* Unsurprisingly, a volume of Planck dimensions cannot contain a mass larger than the Planck mass.

Interesting things happen when we start to determine the error ΔM of a mass measurement in a Planck volume. Let us return to the mass measurement by an orbiting probe. From the relation $GM = rv^2$ we deduce by differentiation that $G\Delta M = v^2 \Delta r + 2vr\Delta v > 2vr\Delta v = 2GM\Delta v/v$. For the error Δv in the velocity measurement we have the indeterminacy relation $\Delta v \geqslant \hbar/(m\Delta r) + \hbar/(MR) \geqslant \hbar/(MR)$. Inserting this in the previous inequality, and forgetting again the factor of 2, we find that the mass measurement error

ΔM of a mass M enclosed in a volume of size R is subject to the condition

$$\Delta M \geqslant \frac{\hbar}{cR} \ . \tag{662}$$

Note that for everyday situations, this error is extremely small, and other errors, such as the technical limits of the balance, are much larger.

To check this result, we can explore another situation. We even use relativistic expressions, in order to show that the result does not depend on the details of the situation or the approximations. Imagine having a mass M in a box of size R and weighing the box with a scale. (It is assumed that either the box is massless or that its mass is subtracted by the scale.) The mass error is given by $\Delta M = \Delta E/c^2$, where ΔE is due to the indeterminacy in the kinetic energy of the mass inside the box. Using the expression $E^2 = m^2c^4 + p^2c^2$, we get that $\Delta M \geqslant \Delta p/c$, which again reduces to equation (662). Now that we are sure of the result, let us continue.

From equation (662) we deduce that for a box of Planck dimensions, the mass measurement *error* is given by the Planck mass. But from above we also know that the mass that can be put inside such a box must not be larger than the Planck mass. Therefore, for a box of Planck dimensions, the mass measurement *error* is larger than (or at best equal to) the mass contained in it: $\Delta M \geqslant M_{\mathrm{Pl}}$. In other words, if we build a balance with two boxes of Planck size, one empty and the other full, as shown in Figure 366, nature cannot decide which way the balance should hang! Note that even a repeated or a continuous measurement will not resolve the situation: the balance will only randomly change inclination, staying horizontal on average.

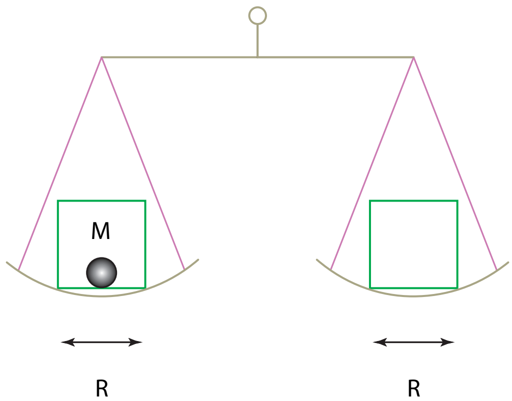


Figure 366 A *Gedanken experiment* showing that at Planck scales, matter and vacuum cannot be distinguished

The argument can be rephrased as follows. The largest mass that we can put in a box of size R is a black hole with a Schwarzschild radius of the same value; the smallest mass present in such a box – corresponding to what we call vacuum – is due to the indeterminacy relation and is given by the mass with a Compton wavelength that matches the size of the box. In other words, inside any box of size R we have a mass m, the limits of which are given by:

$$(\text{full box}) \ \frac{c^2 R}{G} > m > \frac{\hbar}{cR} \ (\text{empty box}) \ . \tag{663}$$

We see directly that for sizes R of the order of the Planck scale, the two limits coincide; in other words, we cannot distinguish a full box from an empty box in that case.

To be sure of this strange result, we check whether it also occurs if, instead of measuring the gravitational mass, as we have just done, we measure the inertial mass. The inertial mass for a small object is determined by touching it, i.e. physically speaking, by performing a scattering experiment. To determine the inertial mass inside a region of size R, a probe must have a wavelength smaller than R, and thus a correspondingly high energy. A high energy means that the probe also attracts the particle through gravity. (We thus find the intermediate result that *at Planck scales, inertial and gravitational mass cannot be distinguished*. Even the balance experiment shown in Figure 366 illustrates this: at Planck scales, the two types of mass are always inextricably linked.) Now, in any scattering experiment, e.g. in a Compton-type experiment, the mass measurement is performed by measuring the wavelength change $\delta\lambda$ of the probe before and after the scattering experiment. The mass indeterminacy is given by

$$\frac{\Delta M}{M} = \frac{\Delta\delta\lambda}{\delta\lambda} . \tag{664}$$

In order to determine the mass in a Planck volume, the probe has to have a wavelength of the Planck length. But we know from above that there always is a minimum wavelength indeterminacy, given by the Planck length l_{Pl}. In other words, for a Planck volume the mass error is always as large as the Planck mass itself: $\Delta M \geqslant M_{\text{Pl}}$. Again, this limit is a direct consequence of the limit on length and space measurements.

This result has an astonishing consequence. In these examples, the measurement error is independent of the mass of the scatterer, i.e. independent of whether or not we start with a situation in which there is a particle in the original volume. We thus find that in a volume of Planck size, it is impossible to say whether or not there is something there when we probe it with a beam!

In short, all arguments lead to the same conclusion: *vacuum, i.e. empty space-time, cannot be distinguished from matter at Planck scales.* Another, often used way to express this state of affairs is to say that when a particle of Planck energy travels through space it will be scattered by the fluctuations of space-time itself, thus making it impossible to say whether it was scattered by empty space-time or by matter. These surprising results rely on a simple fact: whatever definition of mass we use, it is always measured via combined length and time measurements. (This is even the case for normal weighing scales: mass is measured by the displacement of some part of the machine.) The error in these measurements makes it *impossible to distinguish vacuum from matter.*

We can put this result in another way. If on one hand, we measure the mass of a piece of vacuum of size R, the result is always at least $\hbar/cR$; there is no possible way to find a perfect vacuum in an experiment. On the other hand, if we measure the mass of a particle, we find that the result is size dependent; at Planck dimensions it approaches the Planck mass for every type of particle, be it matter or radiation.

If we use another image, when two particles approach each other to a separation of the order of the Planck length, the indeterminacy in the length measurements makes it impossible to say whether there is something or nothing between the two objects. In short, *matter and vacuum are interchangeable at Planck dimensions.* This is an important result: since both mass and empty space-time cannot be differentiated, we have confirmed that they are made of the same 'fabric'. This approach, already suggested above, is now commonplace in all attempts to find a unified description of nature.

This approach is corroborated by the attempts to apply quantum mechanics in highly curved space-time, where a clear distinction between vacuum and particles is impossible.

Ref. 1010 This has already been shown by Fulling–Davies–Unruh radiation. Any accelerated observer and any observer in a gravitational field detects particles hitting him, even if he is Page 800 in vacuum. The effect shows that for curved space-time the idea of vacuum as a particle-free space does not work. Since at Planck scales it is impossible to say whether space is flat or not, it again follows that it is impossible to say whether it contains particles or not.

Curiosities and fun challenges

The strange results at Planck scales imply many other consequences.

▪ The Planck energy is rather large. Imagine that we want to impart this amount of energy to protons using a particle accelerator. How large would that accelerator have to be? Challenge 1374 n In contrast, in everyday life, the Planck energy is rather small. Measured in litres of gasoline, how much fuel does it correspond to? Challenge 1375 n

▪ The usual concepts of matter and of radiation are not applicable at Planck dimensions. Usually, it is assumed that matter and radiation are made up of interacting elementary particles. The concept of an elementary particle is one of an entity that is countable, point-like, real and not virtual, that has a definite mass and a definite spin, that is distinct from its antiparticle, and, most of all, that is distinct from vacuum, which is assumed to have zero mass. All these properties are found to be incorrect at Planck scales. *At Planck dimensions, it does not make sense to use the concepts of 'mass', 'vacuum', 'elementary particle', 'radiation' and 'matter'.*

▪ Do the large mass measurement errors make it possible to claim that mass can be negative at Planck energy? Challenge 1376 n

▪ We now have a new answer to the old question: why is there anything rather than nothing? Well, we now see that at Planck scales there is *no difference* between anything and nothing. In addition, we now can honestly say about ourselves that we are made of nothing.

▪ If vacuum and matter or radiation cannot be distinguished, then it is incorrect to claim that the universe appeared from nothing. The impossibility of making this distinction thus shows that naive creation is a logical impossibility. Creation is not a description of reality. The term 'creation' turns out to be a result of lack of imagination.

▪ Special relativity implies that no length or energy can be invariant. Since we have come to the conclusion that the Planck energy and the Planck length are invariant, there must be deviations from Lorentz invariance at high energy. Can you imagine what the effects would be? Challenge 1377 r In what experiment could they be measured? If you find an answer, publish it; you might get known. Ref. 1012 First attempts are appearing in the research papers. We return to the issue in the third part, with some interesting insights.

▪ Quantum mechanics alone gives, via the Heisenberg indeterminacy relation, a lower limit on the spread of measurements, but strangely enough not on their precision, i.e. not on the number of significant digits. Ref. 1011 Jauch gives the example that atomic lattice constants are known much more precisely than the position indeterminacy of single atoms inside the crystal. Ref. 993

It is sometimes claimed that measurement indeterminacies smaller than the Planck values are possible for large enough numbers of particles. Can you show why this is incorrect, at least for space and time? Challenge 1378 ny

▪ Of course, the idea that vacuum is not empty is not new. More than two thousand years ago, Aristotle argued for a filled vacuum, even though he used incorrect arguments, as seen from today's perspective. In the fourteenth century the discussion on whether empty space was composed of indivisible entities was rather common, but died down again later.

▪ A Planck energy particle falling in a gravitational field would gain energy. However, this is impossible, as the Planck energy is the highest energy in nature. What does this
Challenge 1379 n
imply for this situation?

▪ One way to generalize the results presented here is to assume that, at Planck energy,
Ref. 982
nature is *event symmetric*, i.e. nature is symmetric under exchange of any two events. This idea, developed by Phil Gibbs, provides an additional formulation of the strange behaviour of nature at extreme scales.

Page 452
▪ Because there is a minimum length in nature, so-called naked singularities do not exist. The issue, so hotly debated in the twentieth century, becomes uninteresting, thus ending decades of speculation.

▪ Since mass density and thus energy density are limited, we know that the number of
Page 811
degrees of freedom of any object of finite volume is *finite*. The entropy of black holes has shown us already that entropy values are always finite. This implies that perfect baths do not exist. Baths play an important role in thermodynamics (which is thus found to be only an approximation) and also in recording and measuring devices: when a device measures, it switches from a neutral state to a state in which it shows the result of the measurement.
Ref. 1013
In order to avoid the device returning to the neutral state, it must be coupled to a bath. Without a bath, a reliable measuring device cannot be made. In short, perfect clocks and length measuring devices do not exist because nature puts a limit on their storage ability.

▪ If vacuum and matter cannot be distinguished, we cannot distinguish between ob-
Page 37
jects and their environment. However, this was one the starting points of our journey. Some interesting adventures thus still await us!

▪ We have seen earlier that characterizing nature as made up of particles and vacuum creates problems when interactions are included, since on one hand interactions are the difference between the parts and the whole, while on the other hand, according to
Page 628
quantum theory, interactions are exchanges of particles. This apparent contradiction can be used to show either that vacuum and particles are not the only components of nature, or that something is counted twice. However, since matter and space-time are both made of the same 'stuff,' the contradiction is resolved.

Challenge 1380 d
▪ Is there a smallest possible momentum? And a smallest momentum error?

▪ There is a maximum acceleration in nature. Can you deduce the value of this so-
Challenge 1381 n
called Planck acceleration? Does it require quantum theory?

▪ Given that time becomes an approximation at Planck scales, can we still say whether nature is deterministic? Let us go back to the beginning. We can define time, because in nature change is not random, but gradual. What is the situation now that we know that time is only approximate? Is non-gradual change possible? Is energy conserved? In other words, are surprises possible?

To say that time is not defined at Planck scales and that therefore determinism is an undefinable concept is correct, but not a satisfying answer. What happens at daily life scales? The first answer is that at our everyday scales, the probability of surprises is so small that the world indeed is effectively deterministic. The second answer is that nature is not deterministic, but that the difference is not measurable, since every measurement and observation, by definition, *implies* a deterministic world. The lack of surprises would be due to the limitations of our human nature, and more precisely to the limitations of our senses and brain. The third answer is that the lack of surprises is only apparent, and that we have not yet experienced them yet.

Challenge 1382 n Can you imagine any other possibility? To be honest, it is not possible to answer at this point. But we need to keep the alternatives in mind. We have to continue searching, but with every step we take, we have to consider carefully what we are doing.

▪ If matter and vacuum cannot be distinguished, matter and vacuum each has the properties of the other. For example, since space-time is an extended entity, matter and radiation are also extended entities. Furthermore, as space-time is an entity that reaches the borders of the system under scrutiny, particles must also do so. This is the first hint at the extension of matter; in the following, we will examine this argument in more detail.

▪ Vacuum has zero mass density at large scales, but Planck mass density at Planck scales. Cosmological measurements show that the cosmos is flat or almost flat at large scales, i.e. its energy density is quite low. In contrast, quantum field theory maintains that vacuum has a high energy density (or mass density) at small scales. Since mass is scale dependent, both viewpoints are right, providing a hint to the solution of what is usually called the cosmological constant problem. The contradiction is only apparent; more about this issue later on. Ref. 1014

Challenge 1383 n ▪ When *can* matter and vacuum be distinguished? At what energy?

▪ If matter and vacuum cannot be distinguished, there is a lack of information, which in turn produces an intrinsic basic entropy associated with any part of the universe. We will come back to this topic shortly, in the discussion of the entropy of black holes.

▪ Can we distinguish between liquids and gases by looking at a single atom? No, only by looking at many. In the same way, we cannot distinguish between matter and vacuum by looking at one point, but only by looking at many. We must always *average*. However, even averaging is not completely successful. Distinguishing matter from vacuum is like distinguishing clouds from the clear sky; like clouds, matter also has no defined boundary.

▪ In our exploration we have found that there is no argument which shows that space and time are either continuous or made up of points. Indeed, in contrast, we have found that the combination of relativity and quantum theory makes this impossible. In order to proceed in our ascent of Motion Mountain, we need to leave behind us the usual concept of space-time. *At Planck dimensions, the concept of 'space-time point' or 'mass point' is not applicable in the description of nature.*

Farewell to the big bang

A minimum length, or equivalently,* a minimum action, both imply that there is a maximum curvature for space-time. Curvature can be measured in several ways; for example, surface curvature is an inverse area. A minimum length thus implies a maximum curvature. Within a factor of order one, we find

$$K < \frac{c^3}{G\hbar} = 0.39 \cdot 10^{70}\ \mathrm{m}^{-2}\ . \qquad (665)$$

* The big bang section was added in summer 2002.

as limit for surface curvature K in nature. In other words, the universe never has been a point, never had zero age, never had infinite density and never had infinite curvature. It
Challenge 1384 n
is not difficult to get a similar limit for temperature or any other physical quantity.

In short, since events do not exist, also the big bang cannot have been an event. There never was an initial singularity or a beginning of the universe.

The baggage left behind

In this rapid journey, we have destroyed all the experimental pillars of quantum theory: the superposition principle, space-time symmetry, gauge symmetry, renormalization symmetry and permutation symmetry. We also have destroyed the foundations of general relativity, namely the existence of the space-time manifold, the field concept, the particle concept and the concept of mass. We have even seen that matter and space-time cannot be distinguished.

All these conclusions can be drawn in a simpler manner, by using the minimum action
Page 985
of quantum theory and the maximum force of general relativity. All the mentioned results above are confirmed. It seems that we have lost every concept used for the description of motion, and thus made its description impossible. We naturally ask whether we can save the situation.

First of all, since matter is not distinguishable from vacuum, and since this is true for all types of particles, be they matter or radiation, we have an argument which demonstrates that the quest for unification in the description of elementary particles is correct and necessary.

Moreover, since the concepts 'mass', 'time' and 'space' cannot be distinguished from each other, we also know that a new, *single* entity is necessary to define both particles and space-time. To find out more about this new entity, three approaches are being pursued at the beginning of the twenty-first century. The first, quantum gravity, especially the ap-
Ref. 967
proach using the loop representation and Ashtekar's new variables, starts by generalizing
Ref. 968
space-time symmetry. The second, string theory, starts by generalizing *gauge symmetries* and interactions, while the third, the algebraic quantum group approach, looks for gen-
Ref. 969
eralized *permutation symmetries*. We will describe these approaches in more detail later on.

Before we go on however, we should check with experiments what we have deduced so far.

Some experimental predictions

At present, there is a race going on both in experimental and in theoretical physics: which will be the first experiment that will detect quantum gravity effects, i.e. effects sensitive to the Planck energy?*

One might think that the fluctuations of space and time might make images from far away galaxies unsharp or destroy the phase relation between the photons. However, this
Ref. 1015
effect has been shown to be unmeasurable in all possible cases.

* As more candidates appear, they will be added to this section.

A better candidate is measurement of the speed of light at different frequencies in far away light flashes. There are natural flashes, called gamma ray bursts, which have an extremely broad spectrum, from 100 GeV down to visible light of about 1 eV. These flashes often originate at cosmological distances d. From the difference in arrival time Δt for two frequencies we can define a characteristic energy by setting

$$E_{\text{char}} = \frac{\hbar(\omega_1 - \omega_2)d}{c\Delta t} . \tag{666}$$

Ref. 1016 This energy value is $8 \cdot 10^{16}$ GeV for the best measurement to date. This value is not far from the Planck energy; in fact, it is even closer when the missing factors of order unity are included. It is expected that the Planck scale will be reached in a few years, so that tests will become possible on whether the quantum nature of space-time influences the dispersion of light signals. Planck scale effects should produce a minimum dispersion, different from zero. Detecting it would confirm that Lorentz symmetry is not valid at Planck scales. Ref. 1017, Ref. 1018

Another candidate experiment is the direct detection of distance fluctuations between bodies. Gravitational wave detectors are sensitive to extremely small noise signals in length measurements. There should be a noise signal due to the distance fluctuations induced near Planck energies. The length indeterminacy with which a length l can be measured is predicted to be Ref. 1019, Ref. 1018, Ref. 1020

$$\frac{\delta l}{l} \geqslant \left(\frac{l_{\text{Pl}}}{l}\right)^{2/3} \tag{667}$$

The expression is deduced simply by combining the measurement limit of a ruler in quantum theory with the requirement that the ruler cannot be a black hole. We will discuss this result in more detail in the next section. The sensitivity to noise of the detectors might reach the required level in the early twenty-first century. The noise induced by quantum gravity effects is also predicted to lead to detectable quantum decoherence and vacuum fluctuations. Page 956 Ref. 1021

A third candidate for measurable quantum gravity is the detection of the loss of CPT symmetry at high energies. Especially in the case of the decay of certain elementary particles, such as neutral kaons, the precision of experimental measurement is approaching the detection of Planck scale effects. Ref. 1018

A fourth candidate is the possibility that quantum gravity effects may change the threshold energy at which certain particle reactions become possible. It may be that extremely high energy photons or cosmic rays will make it possible to prove that Lorentz invariance is indeed broken near the Planck scale. Ref. 1022

A fifth candidate is the possibility that the phase of light that travels over long distances gets washed out. However, the first tests show that this is not the case; light form extremely distant galaxies still interferes. The precise prediction of the phase washing effect is still in discussion; most probably the effect is too small to be measured. Ref. 1023

In the domain of atomic physics, it has also been predicted that quantum gravity effects will induce a gravitational Stark effect and a gravitational Lamb shift in atomic transitions. Either effect could be measurable. Ref. 1021

A few candidates for quantum gravity effects have also been predicted by the author.

To get an overview, we summarize and clarify the results found so far.* Special relativity starts with the discovery that observable speeds are limited by the speed of light c. Quantum theory starts with the result that observable actions are limited by $\hbar/2$. Gravitation shows that for every system with length L and mass M, the observable ratio L/M is limited by the constant $4G/c^2$. Combining these results, we have deduced that all physical observables are *bound*, namely by what are usually called the Planck values, though modified by a factor of square root of 2 (or several of them) to compensate for the numerical factors omitted from the previous sentence.

We need to replace $\hbar$ by $\hbar/2$ and G by $4G$ in all the defining expressions for Planck quantities, in order to find the corresponding measurement limits. In particular, the limit for lengths and times is $\sqrt{2}$ times the Planck value and the limit for energy is the Planck value divided by $\sqrt{8}$.* Interestingly, the existence of bounds on all observables makes it possible to deduce several experimentally testable predictions for the unification of quantum theory and general relativity. These predictions do not depend on the detailed final theory.

However, first we need to correct the argument that we have just presented. The argument is only half the story, because we have cheated. The (corrected) Planck values do not seem to be the actual limits to measurements. The actual measurement limits are even stricter still.

First of all, for any measurement, we need certain fundamental conditions to be realized. Take the length measurement of an object. We need to be able to distinguish between matter and radiation, since the object to be measured is made up of matter, and since radiation is the measurement tool that is used to read distances from the ruler. For a measurement process, we need an interaction, which implies the use of radiation. Note that even the use of particle scattering to determine lengths does not invalidate this general requirement.

In addition, for the measurement of wavelengths we need to distinguish between matter and radiation, because matter is necessary to compare two wavelengths. In fact, all
Page 1019 length measurements require the distinction between matter and radiation.** However, this distinction is impossible at the energy of grand unification, when the electroweak
Ref. 1024 and the strong nuclear interactions are unified. At and above this energy, particles of matter and particles of radiation transform into each other; in practice they cannot be distinguished from each other.

If all matter and radiation particles were the same or mixtures of each other, mass could not be defined. Similarly, spin, charge or any other quantum numbers could be defined. To sum up, no measurement can be performed at energies equal to or greater than the GUT unification energy.

In other words, the particle concept (and thus the matter concept) does not run into trouble at the Planck scale, it has already done so at the unification scale. Only below the unification scale do our standard particle and space-time concepts apply. *Only below the unification scale can particles and vacuum be effectively distinguished.*

* This subsection, in contrast to the ones so far, is speculative; it was added in February 2001.

* The entropy of a black hole is thus given by the ratio between its horizon and *half* the minimum area. Of course, a detailed investigation also shows that the Planck mass (divided by $\sqrt{8}$) is the limit for elementary particles from *below* and for black holes from *above*. For everyday systems, there is no limit.

** To speak in modern high energy concepts, all measurements require broken supersymmetry.

As a result, the smallest length in nature is $\sqrt{2}$ times the Planck length reduced by the ratio between the maximal energy $E_{\mathrm{Pl}}/\sqrt{8}$ and the unification energy E_{GUT}. Present
Ref. 1025 estimates give $E_{\mathrm{GUT}} = 10^{16}$ GeV, implying that the smallest accessible length L_{min} is

$$L_{\mathrm{min}} = \sqrt{2}\, l_{\mathrm{Pl}} \frac{E_{\mathrm{Pl}}}{\sqrt{8}\, E_{\mathrm{GUT}}} \approx 10^{-32}\,\mathrm{m} \approx 800\, l_{\mathrm{Pl}} \,. \qquad (668)$$

It is unlikely that measurements at these dimensions will ever be possible. Anyway, the smallest *measurable* length is significantly larger than the Planck scale of nature discussed above. The reason for this is that the Planck scale is that length for which particles and vacuum cannot be distinguished, whereas the minimal measurable length is the distance at which particles of matter and particles of radiation cannot be distinguished. The latter happens at lower energy than the former. We thus have to correct our previous statement to: *the minimum measurable length cannot be smaller than* L_{min}.

The experimentally determined factor of about 800 is one of the great riddles of physics. It is the high-energy equivalent of the quest to understand why the electromagnetic coupling constant is about 1/137, or more simply, why all things have the colours they have. Only the final theory of motion will provide the answer.

In particular, the minimum length puts a bound on the electric dipole moment d of elementary particles, i.e. on any particles without constituents. We get the limit

$$d > d_{\mathrm{min}} = e\, L_{\mathrm{min}} = e\, 10^{-32}\,\mathrm{m} = 1.5 \cdot 10^{-51}\,\mathrm{Cm} \,. \qquad (669)$$

We saw that this result is in contradiction with one of the predictions deduced from the
Page 934 standard model, but not with others. More interestingly, the prediction is in the reach of
Ref. 1026 future experiments. This improved limit may be the simplest possible measurement of yet unpredicted quantum gravity effects. Measuring the dipole moment could thus be a way to determine the unification energy (the factor 800) independently of high-energy physics experiments and possibly to a higher precision.

Interestingly, the bound on the measurability of observables also puts a bound on the measurement *precision* for each observable. This bound is of no importance in everyday life, but it is important at high energy. What is the precision with which a coupling constant can be measured? We can illustrate this by taking the electromagnetic coupling constant as an example. This constant α, also called the fine structure constant, is related to the charge q by

$$q = \sqrt{4\pi\varepsilon_0 \hbar c \alpha} \,. \qquad (670)$$

Now, any electrical charge itself is defined and measured by comparing, in an electrical field, the acceleration to which the charged object is subjected with the acceleration of some unit charge q_{unit}. In other words, we have

$$\frac{q}{q_{\mathrm{unit}}} = \frac{m a}{m_{\mathrm{unit}} a_{\mathrm{unit}}} \,. \qquad (671)$$

Therefore any error in mass and acceleration measurements implies errors in measure-

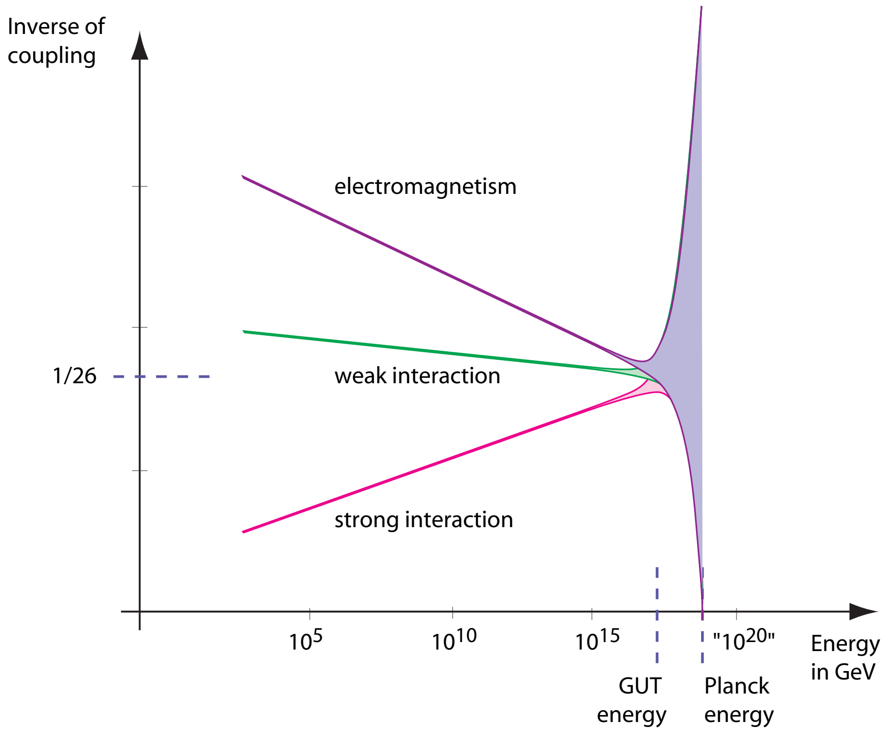


Figure 367 Coupling constants and their spread as a function of energy

ments of charge and the coupling constant.

Page 802 We found in the part on quantum theory that the electromagnetic, the weak and the strong interactions are characterized by coupling constants, the inverse of which depend linearly on the logarithm of the energy. It is usually assumed that these three lines meet at the unification energy already mentioned. Measurements put the unification coupling value at about 1/26. Ref. 1025

We know from the above discussions that the minimum measurement error for any energy measurement at high energies is given by the ratio between the energy to be measured and the limit energy. Inserting this into the graph of the coupling constants 'running' with energy – as physicist like to say – we get the result shown in Figure 367. The search for the consequences of this *fan-out effect* is delightful. One way to put the result is to say that coupling constants are by definition subject to an error. However, all measurement devices, be they clocks, meter rules, scales or any other device, use electromagnetic effects at energies of around 1 eV plus the electron rest energy. This is about 10^{-16} times the GUT energy. As a consequence, the measurement precision of any observable is limited to about 19 digits. The maximum precision presently achieved is 15 digits, and, for the electromagnetic coupling constant, about 9 digits. It will thus be quite some time before this prediction can be tested.

The fun is thus to find a system in which the spreading coupling constant value appears more clearly in the measurements. For example, it may be that high-precision measurements of the g-factor of elementary particles or of high-energy cosmic ray reactions will show some effects of the fan-out. The lifetime of elementary particles could also be affected. Can you find another effect?

Challenge 1385 n

In summary, the experimental detection of quantum gravity effects should be possible, despite their weakness, at some time during the twenty-first century. The successful detection of any such effect will be one of the highlights of physics, as it will challenge the usual description of space and time even more than general relativity did.

We now know that the fundamental entity describing space-time and matter that we are looking for is not point-like. What does it look like? To get to the top of Motion Mountain as rapidly as possible, we will make use of some explosive to blast away a few disturbing obstacles.

Bibliography

958 Steven Weinberg, The cosmological constant problem, *Reviews of Modern Physics* **61**, pp. 1–23, 1989. Cited on page 918.

959 Steven Weinberg, The quantum theory of fields, Cambridge University Press, volume I, 1995, and II, 1996. Cited on page 918.

960 The difficulties are summarised by B.S. de Witt, Quantum field theory in curved space-time, *Physics Reports* **19**, pp. 295–357, 1975. Cited on page 919.

961 C.W. Misner & K.S. Thorne, J.A. Wheeler, *Gravitation*, Freeman, 1973. Cited on pages 919, 922, and 926.

962 J.A. Wheeler, in Relativity, Groups and Topology, edited by C. DeWitt & B. DeWitt (Gordon and Breach, 1994). See also J.A. Wheeler, Physics at the Planck length, *International Journal of Modern Physics A* **8**, pp. 4013–4018, 1993. Cited on page 919.

963 J.L. Friedman & R.D. Sorkin, Spin 1/2 from gravity, *Physical Review Letters* **44**, pp. 1100–1103, 1980. Cited on page 919.

964 A.P. Balachandran, G. Bimonte, G. Marmo & A. Simoni, Topology change and quantum physics, *Nuclear Physics B* **446**, pp. 299–314, 1995, http://www.arxiv.org/abs/hep-th/9503046. Cited on page 919.

965 Many themes discussed here resulted from an intense exchange with Luca Bombelli, now at the Department of Physics and Astronomy of the University of Mississippi. These pages are a reworked version of those published in French as C. Schiller, Le vide diffère-t-il de la matière? in E. Gunzig & S. Diner, editeurs, *Le vide – Univers du tout et du rien – Des physiciens et des philosophes s'interrogent*, Les Éditions de l'Université de Bruxelles, 1998. An older English version is also available as C. Schiller, Does matter differ from vacuum? http://www.arxiv.org/abs/gr-qc/9610066. Cited on page 920.

966 See for example Richard P. Feynman, R.B. Leighton & M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, 1977. Cited on page 920.

967 A. Ashtekar, Recent mathematical developments in quantum general relativity, preprint found at http://www.arxiv.org/abs/gr-qc/9411055. C. Rovelli, Ashtekar Formulation of general relativity and loop-space non-perturbative quantum gravity: a report, *Classical and Quantum Gravity* **8**, pp. 1613–1675, 1991. Cited on pages 921 and 941.

968 M.B. Green, J.H. Schwarz & E. Witten, Superstring theory, volume 1 and 2 (Cambridge University Press, 1987). G. Veneziano, A stringy nature needs just two constants, *Europhysics Letters* **2**, pp. 199–204, 1986. Cited on pages 921 and 941.

969 S. Majid, Introduction to Braided Geometry and q-Minkowski Space, preprint downloadable at http://www.arxiv.org/abs/hep-th/9410241. See also S. Majid, Beyond supersymmetry and quantum symmetry (an introduction to braided groups and braided matrices), in *Quantum Groups, Integrable Statistical Models and Knot Theory*, edited by M.-L. Ge & H.J. de Vega, World Scientific, 1993, pp. 231–282. Cited on pages 921 and 941.

970 C.J. Isham, Prima facie questions in quantum gravity, http://www.arxiv.org/abs/gr-qc/9310031. Cited on page 921.

971 Steven Weinberg, *Gravitation and Cosmology*, (Wiley, 1972). Cited on pages 922, 925, and 926.

972 The observations of black holes at the centre of galaxies and elsewhere are summarised by R. Blandford & N. Gehrels, Revisiting the black hole, *Physics Today* **52**, June 1999. Cited on page 922.

973 M. Planck, Über irreversible Strahlungsvorgänge, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften zu Berlin, pp. 440–480, 1899. Today it is commonplace to use Dirac's $\hbar$ instead of Planck's h, which Planck originally called b. Cited on page 923.

974 The argument is given e.g. in E.P. Wigner, Relativistic invariance and quantum phenomena, *Reviews of Modern Physics* **29**, pp. 255–258, 1957. Cited on page 923.

975 The starting point for the following arguments is taken from M. Schön, Operative time definition and principal indeterminacy, preprint http://www.arxiv.org/abs/hep-th/9304024, and from T. Padmanabhan, Limitations on the operational definition of space-time events and quantum gravity, *Classical and Quantum Gravity* **4**, pp. L107–L113, 1987; see also Padmanabhan's earlier papers referenced there. Cited on page 923.

976 W. Heisenberg, Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik, *Zeitschrift für Physik* **43**, pp. 172–198, 1927. Cited on page 923.

977 E.H. Kennard, Zur Quantenmechanik einfacher Bewegungstypen, *Zeitschrift für Physik* **44**, pp. 326–352, 1927. Cited on page 923.

978 M.G. Raymer, Uncertainty principle for joint measurement of noncommuting variables, *American Journal of Physics* **62**, pp. 986–993, 1994. Cited on page 923.

979 H. Salecker & E.P. Wigner, Quantum limitations of the measurement of space-time distances, *Physical Review* **109**, pp. 571–577, 1958. Cited on page 924.

980 E.J. Zimmerman, The macroscopic nature of space-time, *American Journal of Physics* **30**, pp. 97–105, 1962. Cited on page 924.

981 S.W. Hawking, Particle creation by black holes, *Communications in Mathematical Physics* **43**, pp. 199–220, 1975; see also S.W. Hawking, Black hole thermodynamics, *Physical Review D*13, pp. 191–197, 1976. Cited on page 924.

982 P. Gibbs, The small structure of space-time: a bibliographical review, unpublished preprint found at http://www.arxiv.org/abs/hep-th/9506171. Cited on pages 924 and 939.

983 C.A. Mead, Possible connection between gravitation and fundamental length, *Physical Review B* **135**, pp. 849–862, 1964. Cited on pages 925, 926, 928, and 948.

984 P.K. Townsend, Small-scale structure of space-time as the origin of the gravitational constant, *Physical Review D* **15**, pp. 2795–2801, 1977. Cited on page 926.

985 M.-T. Jaekel & S. Renaud, Gravitational quantum limit for length measurement, *Physics Letters A* **185**, pp. 143–148, 1994. Cited on page 926.

986 D.V. Ahluwalia, Quantum measurement, gravitation and locality, *Physics Letters B* **339**, pp. 301–303, 1994, also preprint http://www.arxiv.org/abs/gr-qc/9308007. Cited on page

926.

987 L. Garay, Quantum gravity and minimum length, *International Journal of Modern Physics A* **10**, pp. 145–165, 1995, also preprint http://www.arxiv.org/abs/gr-qc/9403008; this paper also includes an extensive bibliography. Cited on page 926.

988 G. Amelino–Camelia, Limits on the measurability of space-time distances in (the semiclassical approximation of) quantum gravity, *Modern Physics Letters A* **9**, pp. 3415–3422, 1994, also preprint http://www.arxiv.org/abs/gr-qc/9603014. See also Y.J. Ng & H. Van Dam, Limit to space-time measurement, *Modern Physics Letters A* **9**, pp. 335–340, 1994. Cited on page 926.

989 The generalised indeterminacy relation is an implicit part of reference Ref. 983, especially on page 852, but the issue is explained rather unclearly. Obviously the author found the result too simple to be mentioned explicitly. Cited on pages 926 and 927.

990 C. Rovelli & L. Smolin, Discreteness of area and volume in quantum gravity, *Nuclear Physics B* **442**, pp. 593–619, 1995. R. Loll, The volume operator in discretized quantum gravity, preprint http://www.arxiv.org/abs/gr-qc/9506014. Cited on page 927.

991 D. Amati, M. Ciafaloni & G. Veneziano, Superstring collisions at Planckian energies, *Physics Letters B* **197**, pp. 81–88, 1987. D.J. Gross & P.F. Mende, The high energy behavior of string scattering amplitudes, *Physics Letters B* **197**, pp. 129–134, 1987. K. Konishi, G. Paffuti & P. Provero, Minimum physical length and the generalised uncertainty principle, *Physics Letters B* **234**, pp. 276–284, 1990. P. Aspinwall, Minimum Distances in Non-Trivial String Target Spaces, *Nuclear Physics B* **431**, pp. 78–96, 1994, also preprint http://www.arxiv.org/abs/hep-th/9404060. Cited on page 927.

992 M. Maggiore, A generalised uncertainty principle in quantum mechanics, *Physics Letters B* **304**, pp. 65–69, 1993. Cited on page 927.

993 A simple approach is S. Doplicher, K. Fredenhagen & J.E. Roberts, Space-time quantization induced by classical gravity, *Physics Letters B* **331**, pp. 39–44, 1994, Cited on pages 927 and 938.

994 A. Kempf, Uncertainty relation in quantum mechanics with quantum group symmetry, *Journal of Mathematical Physics* **35**, pp. 4483–4496, 1994. A. Kempf, Quantum Groups and quantum field theory with Nonzero minimal uncertainties in positions and momenta, *Czechoslovak Journal of Physics* **44**, pp. 1041–1048, 1994. Cited on page 927.

995 E.J. Hellund & K. Tanaka, Quantized space-time, *Physical Review* **94**, pp. 192–195, 1954. Cited on page 927.

996 The impossibility to determine temporal ordering in quantum theory is discussed by J. Oppenheimer, B. Reznik & W.G. Unruh, Temporal ordering in quantum mechanics, *Journal of Physics A* **35**, pp. 7641–7652, 2001, or http://www.arxiv.org/abs/quant-ph/0003130. Cited on page 928.

997 A. Peres & N. Rosen, Quantum Limitations on the measurement of gravitational fields, *Physical Review* **118**, pp. 335–336, 1960. Cited on page 928.

998 It is the first definition in Euclid's *Elements*, *c.* 300 bce. For an English translation see T. Heath, *The Thirteen Books of the Elements*, Dover, 1969. Cited on page 929.

999 A beautiful description of the Banach–Tarski paradox is the one by Ian Stewart, Paradox of the spheres, *New Scientist*, 14 January 1995, pp. 28–31. Cited on page 929.

1000 H.S. Snyder, Quantized space-time, *Physical Review* **71**, pp. 38–41, 1947. H.S. Snyder, The electromagnetic field in quantised of space-time, *Physical Review* **72**, pp. 68–74, 1947. A. Schild, Discrete space-time and integral Lorentz transformations, *Physical Review*

73, pp. 414–415, 1948. E.L. Hill, Relativistic theory of discrete momentum space and discrete space-time, *Physical Review* **100**, pp. 1780–1783, 1950. H.T. Flint, The quantization of space-time, *Physical Review* **74**, pp. 209–210, 1948. A. Das, Cellular space-time and quantum field theory, *Il Nuovo Cimento* **18**, pp. 482–504, 1960. Cited on page 931.

1001 D. Finkelstein, 'Superconducting' causal nets, *International Journal of Theoretical Physics* **27**, pp. 473–519, 1985. Cited on page 931.

1002 N.H. Christ, R. Friedberg & T.D. Lee, Random lattice field theory: general formulation, *Nuclear Physics B* **202**, pp. 89–125, 1982. G. 't Hooft, Quantum field theory for elementary particles – is quantum field theory a theory?, *Physics Reports* **104**, pp. 129–142, 1984. Cited on page 931.

1003 See for example L. Bombelli, J. Lee, D. Meyer & R.D. Sorkin, Space-time as a causal set, *Physical Review Letters* **59**, pp. 521–524, 1987. G. Brightwell & R. Gregory, Structure of random space-time, *Physical Review Letters* **66**, pp. 260–263, 1991. Cited on page 931.

1004 For a discussion, see R. Sorabji, *Time, Creation and the continuum: theories in antiquity and the early Middle Ages*, Duckworth, London 1983. Cited on page 931.

1005 For quantum gravity, see for example A. Ashtekar, Quantum geometry in action: big bang and black holes, http://www.arxiv.org/abs/math-ph/0202008. Cited on page 932.

1006 A.D. Sakharov, Vacuum quantum fluctuations in curved space and the theory of gravitation, Soviet Physics – Doklady, **12**, pp. 1040–1041, 1968. Cited on page 934.

1007 C. Wolf, Upper limit for the mass of an elementary particle due to discrete time quantum mechanics, *Il Nuovo Cimento B* **109**, pp. 213–218, 1994. Cited on page 934.

1008 N.F. Ramsey & A. Weis, Suche nach permanenten elektrischen Dipolmomenten: ein Test der Zeitumkehrinvarianz, *Physikalische Blätter* **52**, pp. 859–863, 1996. The paper by E.D. Commins, S.B. Ross, D. DeMille & B.C. Regan, Improved experimental limit on the electric dipole moment of the electron, *Physical Review A* **50**, p. 2960, 1994, gives an upper experimental limit to dipole moment of the electron of $3.4 \cdot 10^{-29}$ e m. W. Bernreuther & M. Suzuki, The electric dipole moment of the electron, *Reviews of Modern Physics* **63**, pp. 313–340, 1991. See also the musings in Hans Dehmelt, Is the electron a composite particle?, *Hyperfine Interactions* **81**, pp. 1–3, 1993. Cited on page 934.

1009 K. Akama, T. Hattori & K. Katsuura, Naturalness bounds on dipole moments from new physics, http://www.arxiv.org/abs/hep-ph/0111238 Cited on page 934.

1010 W.G. Unruh, Notes on black hole evaporation, *Physical Review D* **14**, pp. 870–875, 1976. W.G. Unruh & R.M. Wald, What happens when an accelerating observer detects a Rindler particle, *Physical Review D* **29**, pp. 1047–1056, 1984. Cited on page 938.

1011 W. Jauch, Heisenberg's uncertainty relation and thermal vibrations in crystals, *American Journal of Physics* **61**, pp. 929–932, 1993. Cited on page 938.

1012 One way out is proposed in J. Magueijo & L. Smolin, Lorentz invariance with an invariant energy scale, *Physical Review Letters* **88**, p. 190403, 26 April 2002. They propose a modification of the mass energy relation of the kind

$$E = \frac{\gamma m c^2}{1 + \frac{\gamma m c^2}{E_{\mathrm{Pl}}}} \quad \text{and} \quad p = \frac{\gamma m v}{1 + \frac{\gamma m c^2}{E_{\mathrm{Pl}}}} \, . \tag{672}$$

Another, similar approach of recent years, with a different proposal, is called doubly special relativity. A recent summary is G. Amelino–Camelia, Doubly-special relativity: first results and key open problems, http://www.arxiv.org/abs/gr-qc/0210063. The paper shows how conceptual problems hinder the advance of the field. Cited on page 938.

1013 H.D. Zeh, On the interpretation of measurement in quantum theory, *Foundations of Physics* **1**, pp. 69–76, 1970. Cited on page 939.

1014 Steven Weinberg, The cosmological constant problem, *Reviews of Modern Physics* **61**, pp. 1–23, 1989. Cited on page 940.

1015 See Y.J. Ng, W.A. Christiansen & H. van Dam, Probing Planck-scale physics with extragalactic sources?, *Astrophysical Journal* **591**, pp. L87–L90, 2003, also downloadable electronically at http://www.arxiv.org/abs/astro-ph/0302372; D.H. Coule, Planck scale still safe from stellar images, *Classical and Quantum Gravity* **20**, pp. 3107–3112, 2003, also available as electronic preprint at http://www.arxiv.org/abs/astro-ph/0302333. Negative experimental results (and not always correct calculations) are found in R. Lieu & L. Hillman, The phase coherence of light from extragalactic sources - direct evidence against first order Planck scale fluctuations in time and space, *Astrophysical Journal* **585**, pp. L77–L80, 2003, and R. Ragazzoni, M. Turatto & W. Gaessler, The lack of observational evidence for the quantum structure of spacetime at Planck scales, *Astrophysical Journal* **587**, pp. L1–L4, 2003. Cited on page 941.

1016 B.E. Schaefer, Severe limits on variations of the speed of light with frequency, *Physical Review Letters* **82**, pp. 4964–4966, 21 June 1999. Cited on page 942.

1017 G. Amelino–Camelia, J. Ellis, N.E. Mavromatos, D.V. Nanopoulos & S. Sakar, Potential sensitivity of gamma-ray-burster observations to wave dispersion in vacuo, *Nature* **393**, p. 763, 1998, and preprint http://www.arxiv.org/abs/astro-ph/9712103. Cited on page 942.

1018 G. Amelino–Camelia, Phenomenological description of space-time foam, preprint found at http://www.arxiv.org/abs/gr-qc/0104005, 2001. The paper includes a clearly written overview of present experimental approaches to detecting quantum gravity effects. See also his update G. Amelino–Camelia, Quantum-gravity phenomenology: status and prospects, preprint found at http://www.arxiv.org/abs/gr-qc/0204051, 2002. Cited on page 942.

1019 G. Amelino–Camelia, An interferometric gravitational wave detector as a quantum gravity apparatus, *Nature* **398**, p. 216, 1999, and preprint http://www.arxiv.org/abs/gr-qc/9808029. Cited on page 942.

1020 F. Karolyhazy, Gravitation and quantum mechanics of macroscopic objects, *Il Nuovo Cimento* **A42**, pp. 390–402, 1966, Y.J. Ng & H. van Dam, Limit to space-time measurement, *Modern Physics Letters A* **9**, pp. 335–340, 1994, Y.J. Ng & H. Van Dam, *Modern Physics Letters A* Remarks on gravitational sources, **10**, pp. 2801–2808, 1995. The discussion is neatly summarised in Y.J. Ng & H. van Dam, Comment on 'Uncertainty in measurements of distance', http://www.arxiv.org/abs/gr-qc/0209021. See also Y.J. Ng, Spacetime foam, preprint found at http://www.arxiv.org/abs/gr-qc/0201022. Cited on page 942.

1021 L.J. Garay, Spacetime foam as a quantum thermal bath, *Physics Review Letters* **80**, pp. 2508–2511, 1998, also downloadable as http://www.arxiv.org/abs/gr-qc/9801024. Cited on page 942.

1022 G. Amelino–Camelia & T. Piran, Planck scale-deformation of Lorentz symmetry as a solution to the UHECR and the TeV γ paradoxes, preprint http://www.arxiv.org/astro-ph/0008107, 2000. Cited on page 942.

1023 Experimental tests were described by R. Lieu & L.W. Hillman, The phase coherence of light from extragalactic sources: direct evidence against first-order Planck-scale fluctuations in time and space, *Astrophysical Journal* **585**, pp. L77–L80, 2003, and R. Ragazzoni & al., The lack of observational evidence for the quantum structure of space-time at Planck scales, *Astrophysical Journal* **587**, pp. L1–L4, 2003, Discussions on the predicted phase values are found in http://www.arxiv.org/abs/astro-ph/0302372 and http://www.arxiv.org/abs/astro-ph/0302333. Cited on page 942.

1024 At the unification scale, supersymmetry is manifest; all particles can transform into each other. See for example ... Cited on page 943.

1025 See the paper by U. Amaldi, W. de Boer & H. Fürstenau, Comparison of grand unified theories with elektroweak and strong coupling constants measured at LEP, *Physics Letters* **260**, pp. 447–455, 1991. This is the standard reference on the topic. Cited on pages 944 and 945.

1026 M.V. Romalis, W.C. Griffith, J.P. Jacobs & E.N. Fortson, New limit on the permanent electric dipole moment of ^{199}Hg, *Physical Review Letters* **86**, pp. 2505–2508, 19 March 2001, or http://www.arxiv.org/abs/hep-ex/0012001. Their upper limit on the dipole moment of mercury atoms is $2.1 \cdot 10^{30}$ e m. Research for better measurement methods is progressing; a modern proposal is by S.K. Lamoreaux, Solid state systems for electron electric dipole moment and other fundamental measurements, http://www.arxiv.org/abs/nucl-ex/0109014. Cited on page 944.

35. Nature at large scales – is the universe something or nothing?

Die Grenze ist der eigentlich fruchtbare Ort der Erkenntnis.*

Paul Tillich, *Auf der Grenze*.

This strange question is the topic of the current leg of our mountain ascent. In the last section we explored the properties of nature in the vicinity of Planck dimensions; it is equally fascinating to explore the other limit, namely to study the description of motion at large, cosmological scales. As we proceed, many incredible results will appear, and at the end we will discover a surprising answer to the question in the section title.

Ref. 1027

This section is not standard textbook material; a large part of it is original** and thus speculative and open to question. Even though this section aims at explaining in simple words the ongoing research in the domains of quantum gravity and superstring theory, be warned. With every sentence of this section you will find at least one physicist who disagrees!

We have discussed the universe as a whole several times already. In classical physics we enquired about the initial conditions of the universe and whether it is isolated. In the first intermezzo we asked whether the universe is a set or a concept and indeed, whether it exists at all. In general relativity we gave the classical definition of the term, as the sum of all matter and space-time, we studied the expansion of the universe and we asked about its size and topology. In quantum theory we asked whether the universe has a wavefunction, whether it is born from a quantum fluctuation, and whether it allows the number of particles to be defined.

Here we will settle all these issues by combining general relativity and quantum theory at cosmological scales. That will lead us to some of the strangest results we will encounter in our journey.

Cosmological scales

Hic sunt leones.***

The description of motion requires the application of general relativity whenever the scale d of the situation are of the order of the Schwarzschild radius, i.e. whenever

$$d \approx r_{\mathrm{S}} = 2Gm/c^2 . \qquad (673)$$

It is straightforward to confirm that, with the usually quoted mass m and size d of everything visible in the universe, this condition is indeed fulfilled. We do need general relativity and thus curved space-time when talking about the whole of nature.

Challenge 1386 n

* 'The frontier is the really productive place of understanding'. Paul Tillich (1886–1965), German theologian, socialist and philosopher.

** Written between June and December 2000.

*** 'Here are lions.' Written across unknown and dangerous regions on ancient maps.

Similarly, quantum theory is required for the description of motion of an object whenever we approach it within a distance d of the order of the Compton wavelength λ_C, i.e. whenever

$$d \approx \lambda_C = \frac{h}{mc} \ . \tag{674}$$

Obviously, for the total mass of the universe this condition is not fulfilled. However, we are not interested in the motion of the universe itself; we are interested in the motion of its components. In the description of these components, quantum theory is required whenever pair production and annihilation play a role. This is especially the case in the early history of the universe and near the horizon, i.e. for the most distant events that we can observe in space and time. We are thus obliged to include quantum theory in any precise description of the universe.

Since at cosmological scales we need both quantum theory and general relativity, we start our investigation with the study of time, space and mass, by asking at large scales the same questions that we asked above at Planck scales.

Maximum time

Is it possible to measure time intervals of any imaginable size? General relativity shows that in nature there is a maximum time interval, with a value of about fourteen thousand million years or 430 Ps, providing an upper limit to the measurement of time. It is called the 'age' of the universe and has been deduced from two sets of measurements: the expansion of space-time and the age of matter.

We all know of clocks that have been ticking for a long time: the hydrogen atoms in our body. All hydrogen atoms were formed just after the big bang. We can almost say that the electrons in these atoms have been orbiting the nuclei since the dawn of time. In fact, inside the protons in these atoms, the quarks have been moving already a few hundred thousand years longer. Anyway, we thus get a common maximum time limit for any clock made of atoms. Even 'clocks' made of radiation (can you describe one?) yield a similar maximum time. Also the study of the spatial expansion of the universe leads to the same maximum. No real or imaginable clock or measurement device was ticking *before* this maximum time and no clock could provide a record of having done so.

Challenge 1387 n

In summary, it is *not* possible to measure time intervals greater than the maximum one, either by using the history of space-time or by using the history of matter or radiation.* The maximum time is thus rightly called the 'age' of the universe. Of course, all this is not new, although looking at the issue in more detail does provide some surprises.

* This conclusion implies that so-called 'oscillating' universe models, in which it is claimed that 'before' the big bang there were other phenomena, cannot be based on nature or on observations. They are based on beliefs.

Does the universe have a definite age?

> One should never trust a woman who tells one her real age. A woman who would tell one that, would tell one anything.
>
> Oscar Wilde

Asking about the age of the universe may seem a silly question, because we have just discussed it. Furthermore, the value is found in many books and tables, including that of Page 1071 Appendix B, and its precise determination is actually one of the most important quests in modern astrophysics. But is this quest reasonable?

In order to measure the duration of a movement or the age of a system, we need a clock. The clock has to be *independent* of that movement or system and thus has to be *outside* the system. However, there are no clocks outside the universe and, inside it, a clock cannot be independent. In fact we have just seen that inside the universe, no clock can run throughout its *complete* history. Indeed, time can be defined only once it is possible to distinguish between matter and space-time. Once this distinction can be made, only the two possibilities just discussed remain: we can either talk about the age of space-time, as is done in general relativity, by assuming that matter provides suitable and independent clocks; or we can talk about the age of matter, such as stars or galaxies, by assuming that the extension of space-time or some other matter provides a good clock. Both possibilities are being explored experimentally in modern astrophysics; both give the same result of about fourteen thousand million years that was mentioned previously. However, for the universe as a *whole*, an age cannot be defined.

The issue of the starting point of time makes this difficulty even more apparent. We may imagine that going back in time leads to only two possibilities: either the starting instant $t = 0$ is part of time or it is not. (Mathematically, this means that the segment describing time should be either closed or open.) Both cases assume that it is possible to measure arbitrarily small times, but we know from the combination of general relativity and quantum theory that this is *not* the case. In other words, neither possibility is incorrect: the beginning cannot *be* part of time, nor can it *not be* part of it. To this situation there is only one solution: there was no beginning at all.

In other words, the situation is consistently muddled. Neither the age of the universe nor its origin makes sense. What is going wrong? Or, more correctly, *how* are things going wrong? In other words, what happens if instead of jumping directly to the big bang, we *approach* it as closely as possible? The best way to clarify the issue is to ask about the measurement *error* we are making when we say that the universe is fourteen thousand million years old. This turns out to be a fascinating topic.

How precisely can ages be measured?

> No woman should ever be quite accurate about her age. It looks so calculating.
>
> Oscar Wilde

The first way to measure the age of the universe* is to look at clocks in the usual sense of the term, namely at clocks made of *matter*. As explained in the part on quantum theory,

* Note that the age t_0 is not the same as the Hubble time $T = 1/H_0$. The Hubble time is only a computed

Ref. 1029 Salecker and Wigner showed that a clock built to measure a total time T with a precision Δt has a minimum mass m given by

$$m > \frac{\hbar}{c^2} \frac{T}{(\Delta t)^2} . \tag{675}$$

A simple way to incorporate general relativity into this result was suggested by Ng and
Ref. 1030 van Dam. Any clock of mass m has a minimum resolution Δt due to the curvature of space that it introduces, given by

$$\Delta t > \frac{Gm}{c^3} . \tag{676}$$

If m is eliminated, these two results imply that any clock with a precision Δt can only measure times T up to a certain maximum value, namely

$$T < \frac{(\Delta t)^3}{t_{\mathrm{Pl}}^2} , \tag{677}$$

where $t_{\mathrm{Pl}} = \sqrt{\hbar G/c^5} = 5.4 \cdot 10^{-44}$ s is the already familiar Planck time. (As usual, we have *omitted* factors of order one in this and in all the following results of this section.) In other words, the higher the accuracy of a clock, the shorter the time during which the clock works dependably! The precision of a clock is not (only) limited by the budget spent on building it, but by nature itself. Nevertheless, it does not take much to check that
Challenge 1388 e for clocks in daily life, this limit is not even remotely approached. For example, you may want to deduce how precisely your own age can be specified.

As a consequence of (677), a clock trying to achieve an accuracy of one Planck time can do so for at most one single Planck time! Simply put, *a real clock cannot achieve Planck time accuracy.* If we try to go beyond limit (677), fluctuations of space-time hinder the working of the clock and prevent higher precision. With every Planck time that passes, the clock accumulates a measuring error of at least one Planck time. Thus, the total measurement error is at least as large as the measurement itself. The conclusion is also valid
Challenge 1389 ny for clocks based on radiation, for example on background radiation.

In short, measuring age with a clock always involves some errors; whenever we try to reduce these errors to Planck level, the clock becomes so imprecise that age measurements become impossible.

Does time exist?

Time is waste of money.

Oscar Wilde

From the origins of physics onwards, the concept of 'time' has designated what is meas-
Page 42 ured by a clock. Since equation (677) expresses the non-existence of perfect clocks, it also

quantity and (almost) always larger than the age; the relation between the two depends on the value of the cosmological constant, on the density and on other parameters of the universe. For example, for the standard hot big bang scenario, i.e. for the matter-dominated Einstein–de Sitter model, we have the simple relation
Ref. 1028 $T = (3/2)\, t_0$.

implies that time is only an approximate concept, and that perfect time does not exist. Thus there is no 'idea' of time, in the Platonic sense. In fact, all discussion in the previous and present sections can be seen as proof that there are no perfect or 'ideal' examples of any classical or everyday concept.

Time does not exist. Despite this conclusion, time is obviously a useful concept in everyday life. A simple explanation is provided when we focus on the importance of *energy*. Any clock, in fact any system of nature, is characterized by a simple number, namely the highest ratio of the kinetic energy to the rest energy of its components. In daily life, Challenge 1390 e this fraction is about $1\,\mathrm{eV}/10\,\mathrm{GeV} = 10^{-10}$. Such *low-energy* systems are well suited to building clocks. The more precisely the motion of the main moving part – the pointer of the clock – can be kept constant and can be monitored, the higher the precision of the clock becomes. To achieve the highest possible precision, the highest possible mass of the pointer is required; indeed, both the position and the speed of the pointer must be measured, and the two measurement errors are related by the quantum mechanical indeterminacy relation $\Delta v\,\Delta x > \hbar/m$. High mass implies low intrinsic fluctuations. In order to screen the pointer from outside influences, even more mass is needed. This connection might explain why better clocks are usually more expensive than less accurate ones.

The usually quoted indeterminacy relation is valid only at everyday energies. Increasing the mass does not allow to reach arbitrary small time errors, since general relativity Ref. 1027 changes the indeterminacy relation to $\Delta v\,\Delta x > \hbar/m + G(\Delta v)^2 m/c^3$. The additional term on the right hand side, negligible at everyday scales, is proportional to energy. Increasing it by too large an amount limits the achievable precision of the clock. The smallest Page 926 Challenge 1391 e measurable time interval turns out to be the Planck time.

In summary, time exists as a good approximation only for *low-energy* systems. Any increase in precision beyond a certain limit will require an increase in the energy of the components; at Planck energy, this energy increase will prevent an increase in precision.

What is the error in the measurement of the age of the universe?

Applying the discussion about the measurement of time to the age of the universe is now straightforward. Expression (677) implies that the highest precision possible for a clock Challenge 1392 e is about 10^{-23} s, or about the time light takes to move across a proton. The finite age of the universe also yields a maximum *relative* measurement precision. Expression (677) can be written as

$$\frac{\Delta t}{T} > \left(\frac{t_{\mathrm{Pl}}}{T}\right)^{2/3} \qquad (678)$$

which shows that no time interval can be measured with a precision of more than about 40 decimals.

In order to clarify the issue, we can calculate the error in measurement as a function of the observation energy E_{meas}. There are two limit cases. For *small* energies, the error is given by quantum effects as

$$\frac{\Delta t}{T} \sim \frac{1}{E_{\mathrm{meas}}} \qquad (679)$$

and thus decreases with increasing measurement energy. For high energies, however, the

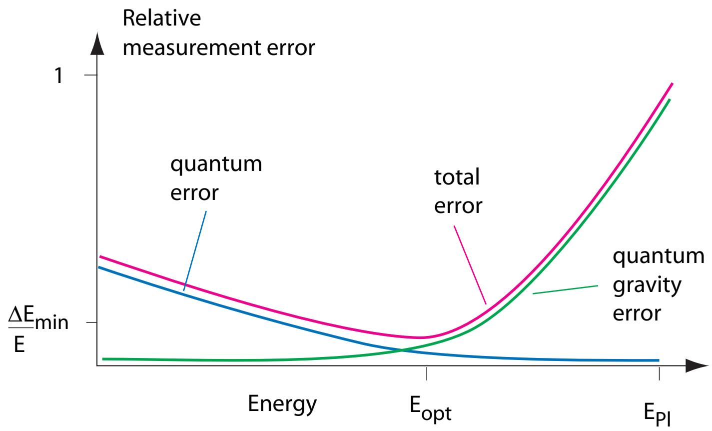


Figure 368 Measurement errors as a function of measurement energy

error is given by gravitational effects as

$$\frac{\Delta t}{T} \sim \frac{E_{\text{meas}}}{E_{\text{Pl}}} \qquad (680)$$

so that the total result is as shown in Figure 368. In particular, energies that are too high do not reduce measurement errors, because any attempt to reduce the measurement error for the age of the universe below 10^{-23} s would require energies so high that the limits of space-time would be reached, making the measurement itself impossible.

We reached this conclusion through an argument based on clocks made of particles. Below we will find out that even by determining the age of the universe using space-time expansion leads to the same limit.

Imagine to observe a tree which, as a result of some storm or strong wind, has fallen towards second tree, touching it at the very top, as shown in Figure 369. It is possible to determine the height of both trees by measuring their separation and the angles at the base. The *error* in height will depend on the errors in separation and angles. Similarly, the age of the universe follows from the present distance and speed of objects – such as galaxies – observed in the night sky. The present distance d corresponds to separation of the trees at ground level and the speed v to the angle between the two trees. The Hubble time T of the universe – which, as has already been mentioned, is usually assumed to be larger than the age of the universe – then corresponds to the height at which the two trees meet. This time since the universe 'started', in a naive sense since the galaxies 'separated', is then given, within a factor of order one, by

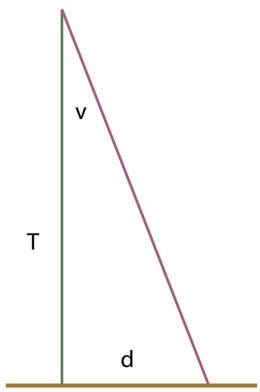


Figure 369 Trees and galaxies

$$T = \frac{d}{v} . \qquad (681)$$

This is in simple words the method used to determine the age of the universe from the expansion of space-time, for galaxies with red-shifts below unity.* Of interest in the following is the (positive) measurement error ΔT, which becomes

$$\frac{\Delta T}{T} = \frac{\Delta d}{d} + \frac{\Delta v}{v} \; . \tag{682}$$

Exploring this in more detail is worthwhile. For any measurement of T we have to choose the object, i.e. a distance d, as well as an observation time Δt, or, equivalently, an observation energy $\Delta E = 2\pi\hbar/\Delta t$. We will now investigate the consequences of these choices for expression (682), always taking into account both quantum theory *and* general relativity.

At everyday energies, the result of the determination of the age t_0 is about $14 \pm 2 \cdot 10^9$ years. This value is deduced by measuring red-shifts, i.e. velocities, and distances, using stars and galaxies in distance ranges from some hundred thousand light years up to a red-shift of about 1. Measuring red-shifts does not produce large velocity errors. The main source of experimental error is the difficulty in determining the distances of galaxies.

What is the smallest possible error in distance? Obviously, equation (678) implies

$$\frac{\Delta d}{T} > \frac{l_{\mathrm{Pl}}^{2/3}}{d^{2/3}} \tag{683}$$

thus giving the same indeterminacy in the age of the universe as found above in the case of material clocks. Challenge 1393 e

We can try to reduce this error in two ways: by choosing objects at either small or large distances. Let us start with the smallest possible distances. In order to get high precision at small distances, we need high observation energies. It is fairly obvious that at observation energies near the Planck value, the value of $\Delta T/T$ approaches unity. In fact, both terms on the right-hand side of expression (682) become of order one. At these energies, Δv approaches c and the maximum value for d approaches the Planck length, for the same reason that at Planck energies the maximum measurable time is the Planck time. In short, *at Planck scales it is impossible to say whether the universe is old or young.* Challenge 1394 e

Let us continue with the other extreme, namely objects extremely far away, say with a red-shift of $z \gg 1$. Relativistic cosmology requires the diagram of Figure 369 to be replaced by the more realistic diagram of Figure 370. The 'light onion' replaces the familiar light cone of special relativity: light converges near the big bang. Ref. 1028

In this case the measurement error for the age of the universe also depends on the distance and velocity errors. At the largest possible distances, the signals an object must send out must be of high energy, because the emitted wavelength must be smaller than the universe itself. Thus, inevitably we reach Planck energies. However, we saw that in such high-energy situations, the emitted radiation, as well as the object itself, are indistinguishable from the space-time background. In other words, the red-shifted signal we

* At higher red-shifts, the speed of light, as well as the details of the expansion, come into play; if we continue with the image of inclined trees, we find that the trees are not straight all the way up to the top and that they grow on a slope, as shown in Figure 370.

would observe today would have a wavelength as large as the size of the universe, with a correspondingly small frequency.

Another way to describe the situation is the following. At Planck energies or near the horizon, the original signal has an error of the same size as the signal itself. When measured at the present time, the red-shifted signal still has an error of the same size as the signal. As a result, the error on the horizon distance becomes as large as the value to be measured.

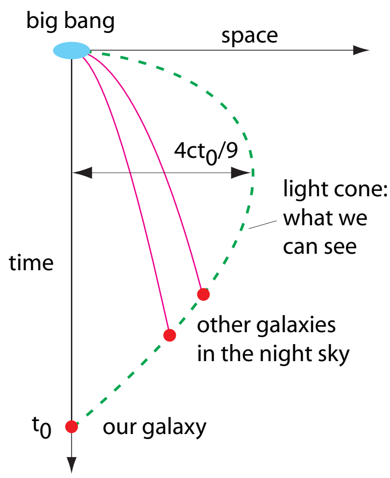


Figure 370 The speed and distance of remote galaxies

In short, even if space-time expansion and large scales are used, the instant of the so-called beginning of the universe cannot be determined with an error smaller than the age of the universe itself, a result we also found at Planck distances. Whenever we aim for perfect precision, we find that the universe is 14 ± 14 thousand million years old! In other words, *at both extremal situations it is impossible to say whether the universe has a non-vanishing age*.

We have to conclude that the anthropomorphic concept of 'age' does not make any sense for the universe as a whole. The usual textbook value is useful only for domains in time, space and energy for which matter and space-time are clearly distinguished, namely at everyday, *human-scale* energies; however, this anthropocentric value has no overall meaning.

Challenge 1395 ny By the way, you may like to examine the issue of the *fate* of the universe using the same arguments. In the text, however, we continue on the path outlined at the start of this section; the next topic is the measurement of length.

Maximum length

General relativity shows that in the standard cosmological model, for hyperbolic (open) and parabolic (marginal) evolutions of the universe, the actual *size* of the universe is infinite. It is only the *horizon distance*, i.e. the distance of objects with infinite red-shift, which is finite. In a hyperbolic or parabolic universe, even though the size is infinite, the most distant visible events (which form the horizon) are at a finite distance.* For elliptical evolution, the total size is finite and depends on the curvature. However, in this case also Page 424 the present measurement limit yields a minimum size for the universe many times larger than the horizon distance. At least, this is what general relativity says.

On the other hand, quantum field theory is based on flat and infinite space-time. Let us see what happens when the two theories are combined. What can we say about measurements of length in this case? For example, would it be possible to construct and use a meter rule to measure lengths larger than the distance to the horizon? It is true that we

* In cosmology, we need to distinguish between the scale factor R, the Hubble radius $c/H = cR/\dot{R}$, the horizon distance h and the size d of the universe. The Hubble radius is a computed quantity giving the distance at which objects move away with the speed of light. It is always *smaller* than the horizon distance, at which in the standard Einstein–de Sitter model, for example, objects move away with *twice* the speed of Ref. 1028 light. However, the horizon itself moves away with *three* times the speed of light.

would have no time to push it up to there, since in the standard Einstein–de Sitter big bang model the horizon moves away from us faster than the speed of light. We should have started using the meter rule right at the big bang.

Ref. 1028

For fun, let us assume that we have actually managed to do this. How far away can we read off distances? In fact, since the universe was smaller in the past and since every observation of the sky is an observation of the past, Figure 370 shows that the maximum *spatial distance* an object can be seen away from us is only $(4/9)ct_0$. Obviously, for space-time intervals, the maximum remains ct_0.

Ref. 1028

Thus, in all cases it turns out to be impossible to measure lengths larger than the horizon distance, even though general relativity predicts such distances. This unsurprising result is in obvious agreement with the existence of a limit for measurements of time intervals. The real surprises come now.

Is the universe really a big place?

Ref. 1031
Page 1071

Astronomers and Hollywood movies answer this question in the affirmative. Indeed, the distance to the horizon of the universe is usually included in tables. Cosmological models specify that the scale factor R, which fixes the distance to the horizon, grows with time t; for the case of the usually assumed mass-dominated Einstein–de Sitter model, i.e. for a vanishing cosmological constant and flat space, we have

$$R(t) = C\, t^{2/3} , \tag{684}$$

where the numerical constant C relates the commonly accepted horizon distance to the commonly accepted age. Indeed, observation shows that the universe is large and is still getting larger. But let us investigate what happens if we add to this result from general relativity the limitations of quantum theory. Is it really possible to measure the distance to the horizon?

We look first at the situation at high energies. We saw above that space-time and matter are not distinguishable at Planck scales. Therefore, at Planck energies we cannot state whether objects are *localized* or not. At Planck scales, a basic distinction of our thinking, namely the one between matter and vacuum, becomes obsolete. Equivalently, it is not possible to claim that space-time is *extended* at Planck scales. Our concept of extension derives from the possibility of measuring distances and time intervals, and from observations such as the ability to align several objects behind one another. Such observations are not possible at Planck scales. In fact, none of the observations in daily life from which we deduce that space is extended are possible at Planck scales. *At Planck scales, the basic distinction between vacuum and matter, namely the opposition between extension and localization, disappears.* As a consequence, at Planck energies the size of the universe cannot be measured. It cannot even be called larger than a match box.

At cosmological distances, the situation is even easier. All the arguments given above on the errors in measurement of the age can be repeated for the distance to the horizon. Essentially, at the largest distances and at Planck energies, the measurement errors are of the same magnitude as the measured value. All this happens because length measurements become impossible at nature's limits. This is corroborated by the lack of any standard with which to compare the size of the universe.

Challenge 1396 ny

Studying the big bang also produces strange results. At Planck energies, whenever we try to determine the size of the big bang, we cannot claim that the universe was smaller than the present size. At Planck energies, there is no way to distinguish length values. Somehow, Planck dimensions and the size of the universe get confused.

There are also other confirmations. Let us go back to the example above. If we had a meter rule spanning the whole universe, even beyond the horizon, with zero at the place where we live, what measurement *error* would it produce for the horizon? It does not take long to discover that the expansion of space-time from Planck scales to the present also expands the indeterminacy in the Planck size into one of the order of the distance to the horizon. The error is as large as the measurement result.

Challenge 1397 ny

Since this also applies when we try to measure the *diameter* of the universe instead of its radius, it becomes impossible to state whether the antipodes in the sky really are distant from each other!

We can summarize the situation by noting that anything said about the size of the universe is as limited as anything said about its age. *The height of the sky depends on the observation energy.* At Planck energies, it cannot be distinguished from the Planck length. If we start measuring the sky at standard observation energies, trying to increase the precision of measurement of the distance to the horizon, the measurement error increases beyond all bounds. At Planck energies, the volume of the universe is indistinguishable from the Planck volume!

The boundary of space-time – is the sky a surface?

The horizon of the universe, essentially the black part of the night sky, is a fascinating entity. Everybody interested in cosmology wants to know what happens there. In newspapers the horizon is sometimes called the *boundary of space-time*. Some surprising insights, not yet common in newspapers, appear when the approaches of general relativity and quantum mechanics are combined.

We saw above that the errors in measuring the distance of the horizon are substantial. They imply that we cannot pretend that all points of the sky are equally far away from us. Thus we cannot say that the sky is a surface. There is even no way to determine the dimensionality of the horizon or the dimensionality of space-time near the horizon.*

Measurements thus do not allow us to determine whether the boundary is a point, a surface, or a line. It may be an arbitrary complex shape, even knotted. In fact, quantum theory tells us that it must be all of these from time to time, in short, that *the sky fluctuates in height and shape*. In short, it is impossible to determine the topology of the sky. But that is nothing new. As is well known, general relativity is unable to describe pair creation particles with spin 1/2. The reason for this is the change in space-time topology required by the process. On the other hand, the universe is full of such processes, implying that it is impossible to define a topology for the universe and, in particular, to talk of the topology of the horizon itself. Are you able to find at least two other arguments to show this?

Challenge 1399 ny

* In addition, the measurement errors imply that no statement can be made about translational symmetry at cosmological scales. Are you able to confirm this? In addition, at the horizon it is impossible to distinguish between spacelike and timelike distances. Even worse, concepts such as 'mass' or 'momentum' are muddled at the horizon. This means that, as at Planck energies, we are unable to distinguish between objects and the background, and between state and intrinsic properties. We will come back to this important point shortly.

Challenge 1398 ny

Worse still, quantum theory shows that space-time is not continuous at a horizon, as
Ref. 1027 can easily be deduced by applying the Planck-scale arguments from the previous section. Time and space are not defined there.

Finally, there is no way to decide whether the various boundary points are *different* from each other. The distance between two points in the night sky is undefined. In other words, it is unclear what the *diameter* of the horizon is.

In summary, the horizon has no specific distance or shape. The horizon, and thus the universe, cannot be shown to be manifolds. This leads to the next question:

Does the universe have initial conditions?

Ref. 1032 One often reads about the quest for the initial conditions of the universe. But before joining this search, we should ask *whether* and *when* such initial conditions make any sense. Obviously, our everyday description of motion requires them. Initial conditions describe the *state* of a system, i.e. all those aspects that differentiate it from a system with the same intrinsic properties. Initial conditions, like the state of a system, are attributed to a system by an *outside* observer.

More specifically, quantum theory tells us that initial conditions or the state of a system can only be defined by an outside observer with respect to an environment. It is already difficult to be outside the universe. In addition, quite independently of this issue, even inside the universe a state can only be defined if matter can be distinguished from vacuum. However, this is impossible at Planck energies, near the big bang, or at the horizon. Thus the universe has no state. No state also means that there is *no wavefunction of the universe*.

Ref. 1027 The limits imposed by the Planck values confirm this conclusion in other ways. First of all, they show that the big bang was not a singularity with infinite curvature, density or temperature, because infinitely large values do not exist in nature. Second, since instants of time do not exist, it is impossible to define the state of any system at a given time. Third, as instants of time do not exist, neither do events exist, and thus the big bang was not an event, so that for this more prosaic reason, neither an initial state nor an initial wavefunction can be ascribed to the universe. (Note that this also means that the universe cannot have been created.)

In short, *there are no initial conditions for the universe*. Initial conditions make sense only for subsystems and only far away from Planck scales. Thus, for initial conditions to exist, the system must be far away from the horizon and it must have evolved for some time 'after' the big bang. Only when these two requirements are fulfilled can objects *move* in space. Of course, this is always the case in everyday life.

At this point in our mountain ascent, where time and length are unclearly defined at cosmological scales, it should come as no surprise that there are similar difficulties concerning the concept of mass.

Does the universe contain particles and stars?

The number of stars, about $10^{23\pm1}$, is included in every book on cosmology, as it is in the table of Appendix B.1071 A subset of this number can be counted on clear nights. If we ask the same question about particles instead of stars, the situation is similar. The commonly quoted number of baryons is $10^{81\pm1}$, together with $10^{90\pm1}$ photons. However, this does not settle the issue. Neither quantum theory nor general relativity alone make predictions about the number of particles, either inside or outside the horizon. What happens if we combine them?

Page 721
In order to define the number of particles in a region, quantum theory first of all requires a vacuum state to be defined. The number of particles is defined by comparing the system with the vacuum. If we neglect or omit general relativity by assuming flat space-time, this procedure poses no problem. However, if we include general relativity and thus a curved space-time, especially one with such a strangely behaved horizon as the one we have just found, the answer is simple: there is *no* vacuum state with which we can compare the universe, for two reasons. First, nobody can explain what an empty universe would look like; second, and more importantly, there is no way to define a state of the universe at all. The number of particles in the universe thus becomes undefinable. Only at everyday energies and for finite dimensions are we able to speak of an approximate number of particles.

Comparison between a system and the vacuum is also impossible in the case of the universe for purely practical reasons. The requirement for such a comparison effectively translates into the requirement that the particle counter be outside the system. (Can you
Challenge 1400 ny
confirm the reason for this connection?) In addition, it is impossible to remove particles from the universe. The impossibility of defining a vacuum state, and thus the number of particles in the universe, is not surprising. It is an interesting exercise to investigate the measurement errors that appear when we try to determine the number of particles
Challenge 1401 ny
despite this fundamental impossibility.

Can we count the stars? In principle, the same conclusion applies as for particles. However, at everyday energies the stars can be counted *classically*, i.e. without taking them out of the volume in which they are enclosed. For example, this is possible if the stars are differentiated by mass, colour or any other individual property. Only near Planck energies or near the horizon are these methods inapplicable. In short, the number of stars is only defined as long as the observation energy is low, i.e. as long as we stay away from Planck energies and from the horizon.

Therefore, despite what appear to be the case on human scales, *there is no definite number of particles in the universe*. The universe cannot be distinguished from vacuum by counting particles. Even though particles are necessary for our own existence and functioning, a complete count of them cannot be made.

This conclusion is so strange that we should not accept it too easily. Let us try another method of determining the content of matter in the universe: instead of counting particles, let us weigh them.

Does the universe contain masses and objects?

Page 1071
The average density of the universe, about 10^{-26} kg/m^3, is frequently cited in texts. Is it different from a vacuum? Quantum theory shows that, as a result of the indeterminacy relation, even an empty volume of size R has a mass. For a zero-energy photon inside such a vacuum, we have $E/c = \Delta p > \hbar/\Delta x$, so that in a volume of size R, we have a minimum mass of at least $m_{min}(R) = h/cR$. For a spherical volume of radius R there is

thus a minimal mass density given approximately by

$$\rho_{\min} \approx \frac{m_{min}(R)}{R^3} = \frac{\hbar}{cR^4} . \qquad (685)$$

For the universe, if the standard horizon distance R_0 of fourteen thousand million light years is inserted, the value becomes about 10^{-142} kg/m^3. This describes the density of the vacuum. In other words, the universe, with a density of about 10^{-26} kg/m^3, seems to be clearly different from vacuum. But are we sure?

We have just deduced that the radius of the horizon is undefined: depending on the observation energy, it can be as small as the Planck length. This implies that the density of the universe lies somewhere between the lowest possible value, given by the density of vacuum just mentioned, and the highest possible one, namely the Planck density.* In short, relation (685) does not really provide a clear statement.

Another way to measure the mass of the universe would be to apply the original definition of mass, as given by Mach and modified by special relativity. Thus, let us try to collide a standard kilogram with the universe. It is not hard to see that whatever we do, using either low or high energies for the standard kilogram, the mass of the universe cannot be constrained by this method. We would need to produce or to measure a velocity change Δv for the rest of the universe after the collision. To hit all the mass in the universe at the same time, we need high energy; but then we are hindered by Planck energy effects. In addition, a properly performed collision measurement would require a mass outside the universe, a rather difficult feat to achieve.

Page 283

Challenge 1403 ny

Yet another way to measure the mass would be to determine the gravitational mass of the universe through straightforward weighing. But the lack of balances outside the universe makes this an impractical solution, to say the least.

A way out might be to use the most precise definition of mass provided by general relativity, the so-called *ADM mass*. However, for definition this requires a specified behaviour at infinity, i.e. a background, which the universe lacks.

We are then left with the other general relativistic method: determining the mass of the universe by measuring its average curvature. Let us take the defining expressions for average curvature κ for a region of size R, namely

Page 382

$$\kappa = \frac{1}{r^2_{\text{curvature}}} = \frac{3}{4\pi}\,\frac{4\pi R^2 - S}{R^4} = \frac{15}{4\pi}\,\frac{4\pi R^3/3 - V}{R^5} . \qquad (687)$$

We have to insert the horizon radius R_0 and either its surface area S_0 or its volume V_0. However, given the error margins on the radius and the volume, especially at Planck energies, for the radius of curvature we again find no reliable result.

Challenge 1404 ny

* In fact, at everyday energies the density of the universe lies almost exactly between the two values, yielding the strange relation

Challenge 1402 ny

$$m_0^2/R_0^2 \approx m_{\text{Pl}}^2/R_{\text{Pl}}^2 = c^4/G^2 . \qquad (686)$$

But this is nothing new. The approximate equality can be deduced from equation 16.4.3 (p. 620) of Steven Weinberg, *Gravitation and Cosmology*, Wiley, 1972, namely $Gn_b m_p = 1/t_0^2$. The relation is required by several cosmological models.

Page 810

An equivalent method starts with the usual expression for the indeterminacy $\Delta\kappa$ in the scalar curvature for a region of size R provided by Rosenfeld, namely

Ref. 1033

$$\Delta\kappa > \frac{16\pi l_{\mathrm{Pl}}^2}{R^4} . \tag{688}$$

However, this expression also shows that the error in the radius of curvature behaves like the error in the distance to the horizon.

In summary, *at Planck energies, the average radius of curvature of nature turns out to lie between infinity and the Planck length.* This implies that the density of matter lies between the minimum value and the Planck value. There is thus no method to determine the mass of the universe at Planck energies. (Can you find one?) The concept of mass cannot be applied to the universe as a whole. *Thus, the universe has no mass.*

Challenge 1405 ny

Do symmetries exist in nature?

We have already seen that at the horizon, space-time translation symmetry breaks down. Let us have a quick look at the other symmetries.

What happens to permutation symmetry? Exchange is an operation on objects in space-time. Exchange thus automatically requires a distinction between matter, space and time. If we cannot distinguish positions, we cannot talk about exchange of particles. However, this is exactly what happens at the horizon. In short, general relativity and quantum theory together make it impossible to define permutation symmetry at the horizon.

CPT symmetry suffers the same fate. As a result of measurement errors or of limiting maximum or minimum values, it is impossible to distinguish between the original and the transformed situations. It is therefore impossible to maintain that CPT is a symmetry of nature at horizon scales. In other words, matter and antimatter cannot be distinguished at the horizon.

Challenge 1406 ny

The same happens with gauge symmetry, as you may wish to check in detail yourself. For its definition, the concept of gauge field requires a distinction between time, space and mass; at the horizon this is impossible. We therefore also deduce that at the horizon also concepts such as algebras of observables cannot be used to describe nature. Renormalization also breaks down.

All symmetries of nature break down at the horizon. The complete vocabulary we use when we talk about observations, including terms such as such as magnetic field, electric field, potential, spin, charge, or speed, cannot be used at the horizon. And that is not all.

Does the universe have a boundary?

It is common to take 'boundary' and 'horizon' as synonyms in the case of the universe, because they are the same for all practical purposes. To study this concept, knowledge of mathematics does not help us; the properties of mathematical boundaries, e.g. that they themselves have no boundary, are not applicable in the case of nature, since space-time is not continuous. We need other, physical arguments.

The boundary of the universe is obviously supposed to represent the boundary between *something* and *nothing*. This gives three possibilities:

- 'Nothing' could mean 'no matter'. But we have just seen that this distinction cannot be made at Planck scales. As a consequence, the boundary will either not exist at all or it will encompass the horizon *as well as* the whole universe.
- 'Nothing' could mean 'no space-time'. We then have to look for those domains where space and time cease to exist. These occur at Planck scales and at the horizon. Again, the boundary will either not exist or it will encompass the whole universe.
- 'Nothing' could mean 'neither space-time nor matter.' The only possibility is a boundary that encloses domains *beyond* the Planck scale and *beyond* the horizon; but again, such a boundary would also encompass all of nature.

This result is puzzling. When combining quantum theory and relativity, we do not seem to be able to find a conceptual definition of the horizon that distinguishes the horizon from what it includes. In fact, if you find one, publish it! A distinction *is* possible in general relativity; and equally, a distinction *is* possible in quantum theory. However, as soon as we combine the two, the boundary becomes indistinguishable from its content. *The interior of the universe cannot be distinguished from its horizon*. There is no boundary.

Challenge 1407 ny

A difficulty to distinguish the horizon and its contents is definitely interesting; it suggests that nature may be *symmetric* under transformations that exchange interiors and boundaries. This connection is nowadays called *holography* because it vaguely recalls the working of credit card holograms. It is a busy research field in present-day high-energy physics. However, for the time being, we shall continue with our original theme, which directly leads us to ask:

Ref. 1034

Is the universe a set?

We are used to calling the universe the sum of all matter and all space-time. In other words, we imply that the universe is a set of components, all different from each other. This idea was introduced in three situations: it was assumed that matter consists of particles, that space-time consists of events (or points) and that the set of states consists of different initial conditions. However, our discussion so far shows that the universe is *not* a set of such distinguishable elements. We have encountered several proofs: at the horizon, at the big bang and at Planck scales distinction between events, between particles, between observables and between space-time and matter becomes impossible. In those domains, distinctions of any kind become impossible. We have found that any distinction between two entities, such as between a toothpick and a mountain, is only *approximately* possible. The approximation is possible because we live at energies much smaller than the Planck energy. Obviously, we are able to distinguish cars from people and from toothpicks; the approximation is so good that we do not notice the error when we perform it. Nevertheless, the discussion of the situation at Planck energies shows that a perfect distinction is impossible in principle. *It is impossible to split the universe into separate entities.*

Another way to reach this result is the following. Distinguishing between two entities requires different measurement results, such as different positions, masses, sizes, etc. Whatever quantity we choose, at Planck energies the distinction becomes impossible. Only at everyday energies is it approximately possible.

In short, since the universe contains no distinguishable entities, *the universe is not a set*. We have already envisaged this possibility in the first intermezzo; now it is confirmed.

Page 633

The concept of 'set' is already too specialized to describe the universe. *The universe must be described by a mathematical concept that does not contain any set.*

This is a powerful result: it means that the universe cannot be described precisely if any of the concepts used for its description presuppose the use of sets. But all concepts we have used so far to describe nature, such as space-time, phase space, Hilbert space and its generalizations, namely Fock space and particle space, are based on sets. They all must be abandoned at Planck energies, as well as in any precise description.

Furthermore, many speculations about unified descriptions do not satisfy the criterion that sets must not be included. In particular, all studies of quantum fluctuations, mathematical categories, posets, complex mathematical spaces, computer programs, Turing machines, Gödel's theorem, creation of any sort, space-time lattices, quantum lattices and Bohm's unbroken wholeness fail to satisfy this requirement. In addition, almost none of the speculations about the origin of the universe can be correct. For example, you may wish to check the religious explanations you know against this result. In fact, no approach used theoretical physics in the year 2000 satisfies the requirement that sets must be abandoned; perhaps a future version of string or M theory will do so. The task is not easy; do you know of a single concept not based on a set?

Challenge 1408 ny

Challenge 1409 n

Note that this radical conclusion is deduced from only two statements: the necessity of using quantum theory whenever the dimensions are of the order of the Compton wavelength, and the necessity to use general relativity whenever the dimensions are of the order of the Schwarzschild radius. Together, they mean that any precise description of nature cannot contain sets. We have reached this result after a long and interesting, but in a sense unnecessary, digression. The difficulties in complying with this result may explain why the unification of the two theories has not so far been successful. Not only does unification require that we stop using space, time and mass for the description of nature; it also requires that all distinctions, of any kind, should be only approximate. But all physicists have been educated on the basis of exactly the opposite creed!

Note that, because it is not a set, *the universe is* not *a physical system.* Specifically, it has no state, no intrinsic properties, no wavefunction, no initial conditions, no density, no entropy and no cosmological constant. The universe is thus neither thermodynamically closed nor open; and it contains no information. All thermodynamic quantities, such as entropy, temperature and free energy, are defined using *ensembles*. Ensembles are limits of systems which are thermodynamically either open or closed. As the universe is neither of these two, no thermodynamic quantity can be defined for it.* All physical properties are defined only for parts of nature which are approximated or idealized as sets, and thus are physical systems.

* Some people knew this long before physicists; for example, the belief that the universe is or contains information was ridiculed most thoroughly in the popular science fiction parody by DOUGLAS ADAMS, *The Hitchhiker's Guide to the Galaxy*, 1979, and its sequels.

Curiosities and fun challenges

> Insofern sich die Sätze der Mathematik auf die Wirklichkeit beziehen, sind sie nicht sicher, und sofern sie sicher sind, beziehen sie sich nicht auf die Wirklichkeit.*
>
> Albert Einstein

The contradictions between the term 'universe' and the concept of 'set' lead to numerous fascinating issues. Here are a few.

▪ In mathematics, 2 + 2 = 4. This statement is an idealization of statements such as 'two apples plus two apples makes four apples.' However, we now know that at Planck energies this is not a correct statement about nature. At Planck energies, objects cannot be counted. Are you able to confirm this? Challenge 1410 ny

▪ In 2002, Seth Lloyd estimated how much information the universe can contain, and how many calculations it has performed since the big bang. This estimate is based on two ideas: that the number of particles in the universe is a well-defined quantity, and that the universe is a computer, i.e. a physical system. We now know that neither assumption is correct. This example shows the power of the criteria that we deduced for the final description of motion. Ref. 1035

▪ People take pictures of the cosmic background radiation and its variations. Is it possible that the photographs will show that the spots in one direction of the sky are exactly the same as those in the diametrically opposite direction? Challenge 1411 ny

▪ In 1714, Leibniz published his *Monadologie*. In it he explores what he calls a simple substance, which he defined to be a substance that has no parts. He called it a *monade* and explores some of its properties. However, due mainly to its incorrect deductions, the term has not been taken over by others. Let us forget the strange deductions and focus only on the definition: what is the physical concept most related to that of monade? Ref. 1036 Challenge 1412 n

▪ We usually speak of *the* universe, implying that there is only one of them. Yet there is a simple case to be made that 'universe' is an observer-dependent concept, since the idea of 'all' is observer-dependent. Does this mean that there are many universes? Challenge 1413 ny

▪ If all particles would be removed – assuming one would know where to put them – there wouldn't be much of a universe left. True? Challenge 1414 ny

▪ At Planck energies, interactions cannot be defined. Therefore, 'existence' cannot be defined. In short, at Planck energies we cannot say whether particles exist. True? Challenge 1415 ny

Hilbert's sixth problem settled

In the year 1900, David Hilbert gave a famous lecture in which he listed 23 of the great challenges facing mathematics in the twentieth century. Most of these problems provided challenges to many mathematicians for decades afterwards. A few are still unsolved, among them the sixth, which challenged mathematicians and physicists to find an *axiomatic* treatment of physics. Page 609 Ref. 1037

* In so far as mathematical statements describe reality, they are not certain, and as far as they are certain, they are not a description of reality.

Since the universe is not even a set, we can deduce that such an axiomatic description of nature is *impossible*. The reasoning is simple; all mathematical systems, be they algebraic systems, order systems or topological systems, are based on sets. Mathematics does not have axiomatic systems that do not contain sets. The reason for this is simple: any (mathematical) concept contains at least one set. However, nature does not.

The impossibility of an axiomatic system for physics is also confirmed in another way. Physics starts with a *circular* definition: space-time is defined with the help of objects and Page 144 objects are defined with the help of space-time. Physics thus has *never* been modelled on the basis of mathematics. Physicists have always had to live with logical problems.

The situation is similar to a child's description of the sky as 'made of air and clouds'. Looking closely, we discover that clouds are made up of water droplets. However, there is air inside clouds, and there is also water vapour elsewhere in the air. When clouds and air are viewed through a microscope, there is no clear boundary between the two. We cannot define either of the terms 'cloud' and 'air' without the other. No axiomatic definition is possible.

Objects and vacuum also behave in this way. Virtual particles are found in vacuum, and vacuum is found inside objects. At Planck scales there is no clear boundary between the two; we cannot define either of them without the other. Despite the lack of a precise definition and despite the logical problems that ensue, in both cases the description works well at large, everyday scales.

We note that, since the universe is not a set and since it contains no information, the Page 241 paradox of the physics book containing a full description of nature disappears. Such a book can exist, as it does not contradict any property of the universe. But then a question arises naturally:

Does the universe make sense?

> Drum hab ich mich der Magie ergeben,
> [...]
> Daß ich erkenne, was die Welt
> Im Innersten zusammenhält.*
>
> Goethe, *Faust*.

Is the universe really the sum of matter–energy and space-time? Or of particles and vacuum? We have heard this so often up to now that we may be lulled into forgetting to check the statement. To find the answer, we do not need magic, as Faust thought; we only need to list what we have found so far, especially in this section, in the section on Planck Page 584 scales, and in the intermezzo on brain and language. Table 71 shows the result.

Not only are we unable to state that the universe is made of space-time and matter; in fact, we are unable to say anything about the universe at all!** It is not even possible to say that it exists, since it is impossible to interact with it. The term 'universe' does not Challenge 1417 r allow us to make a single sensible statement. (Can you find one?) We are only able to say which properties it does *not* have. We are unable to find any property the universe *does*

* Thus I have devoted myself to magic, [...] that I understand how the innermost world is held together.

** There is also another well-known, non-physical concept about which nothing can be said. Many scholars Challenge 1416 n have explored it in detail. Can you see what it is?

Table 71 Physical statements about the universe

▪ The universe has no age.	▪ The universe has no beginning.
▪ The universe has no size.	▪ The universe has no volume.
▪ The universe has no shape.	▪ The universe's particle number is undefined.
▪ The universe has no mass.	▪ The universe has no energy.
▪ The universe has no density.	▪ The universe contains no matter.
▪ The universe has no cosmological constant.	▪ The universe has no initial conditions.
▪ The universe has no state.	▪ The universe has no wave function.
▪ The universe is not a physical system.	▪ The universe contains no information.
▪ The universe is not isolated.	▪ The universe is not open.
▪ The universe has no boundaries.	▪ The universe does not interact.
▪ The universe cannot be measured.	▪ The universe cannot be said to exist.
▪ The universe cannot be distinguished from nothing.	▪ The universe cannot be distinguished from a single event.
▪ The universe contains no moments.	▪ The universe is not composite.
▪ The universe is not a set.	▪ The universe is not a concept.
▪ The universe cannot be described.	▪ There is no plural for 'universe'.
▪ The universe cannot be distinguished from vacuum.	▪ The universe was not created.

have. Thus, the universe has no properties! We cannot even say whether the universe is something or nothing. *The universe isn't anything in particular.* In other words, the term 'universe' is not useful at all for the description of motion.

Page 597 We can obtain a confirmation of this strange conclusion from the first intermezzo. There we found that any concept needs defined content, defined limits and a defined domain of application. In this section, we have found that for the term 'universe', not one of these three aspects is defined; there is thus *no* such concept. If somebody asks: 'why does the universe exist?' the answer is: not only does the use of 'why' wrongly suggest that something may exist outside the universe, providing a reason for it and thus contradicting the definition of the term 'universe' itself; most importantly of all, the universe simply does not exist. In summary, any sentence containing the word universe makes no sense. The term 'universe' only *seems* to express something, even if it doesn't.*

The conclusion that the term universe makes no sense may be interesting, even strangely beautiful; but does it help us to understand motion more precisely? Interestingly so, it does.

A concept without a set eliminates contradictions

The discussion about the term 'universe' also shows that the term does not contain any set. In other words, this is the first term that will help us on the way to a precise description

* Of course, the term 'universe' still makes sense if it is defined more restrictively, such as 'everything interacting with a particular human or animal observer in everyday life'. But such a definition is not useful for our quest, as it lacks the precision required for any description of motion.

of nature. We will see later on how this happens.

By taking into account the limits on length, time, mass and all other quantities we have encountered, we have deduced a number of almost painful conclusions about nature. However, we also received something in exchange: all the contradictions between general relativity and quantum theory that we mentioned at the beginning of this chapter are now resolved. Although we have had to leave behind us many cherished habits, in exchange we have the promise of a description of nature without contradictions. But we get even more.

Challenge 1418 e

Extremal scales and open questions in physics

Page 887

In the chapter *Quantum physics in a nutshell* we listed all the unexplained properties of nature left open either by general relativity or by quantum theory. The present conclusions provide us with new connections among them. Indeed, many of the cosmological results of this section sound surprisingly familiar; let us compare them systematically with those of the section on Planck scales. Both sections explored topics – some in more detail than others – from the list of unexplained properties of nature.

First, Table 72 shows that *none* of the unexplained properties makes sense at *both* limits of nature, the small *and* the large. All open questions are open at both extremes. Second and more importantly, nature behaves in the same way at horizon scales and at Planck scales. In fact, we have not found any difference between the two cases. (Are you able to discover one?) We are thus led to the hypothesis that nature does not distinguish between the large and the small. Nature seems to be characterized by *extremal identity*.

Challenge 1419 r

Is extremal identity a principle of nature?

The principle of extremal identity incorporates some rather general points:

- all open questions about nature appear at its two extremes;
- a description of nature requires both general relativity and quantum theory;
- nature or the universe is not a set;
- initial conditions and evolution equations make no sense at nature's limits;
- there is a relation between local and global issues in nature;
- the concept of 'universe' has no content.

Extremal identity thus looks like a good candidate tool for use in the search for a unified description of nature. To be a bit more provocative, it may be the *only* known principle incorporating the idea that the universe is not a set, and thus might be the only candidate tool for use in the quest of unification. Extremal identity is beautiful in its simplicity, in its unexpectedness and in the richness of its consequences. You might enjoy exploring it by yourself.

Challenge 1420 ny

The study of the consequences of extremal identity is currently the focus of much activity in high energy particle physics, although often under different names. The simplest approach to extremal identity – in fact one that is too simple to be correct – is inversion. It looks as if extremal identity implies a connection such as

Ref. 1038

$$r \leftrightarrow \frac{l_{\mathrm{Pl}}^2}{r} \quad \text{or} \quad x_\mu \leftrightarrow \frac{l_{\mathrm{Pl}}^2\, x_\mu}{x_\mu\, x^\mu} \qquad (689)$$

Table 72 Properties of nature at maximal, everyday and minimal scales

Physical property of nature	At horizon scale	At everyday scale	At Planck scale
requires quantum theory and relativity	true	false	true
intervals can be measured precisely	false	true	false
length and time intervals are	limited	unlimited	limited
space-time is not continuous	true	false	true
points and events cannot be distinguished	true	false	true
space-time is not a manifold	true	false	true
space is 3 dimensional	false	true	false
space and time are indistinguishable	true	false	true
initial conditions make sense	false	true	false
space-time fluctuates	true	false	true
Lorentz and Poincaré symmetry	does not apply	applies	does not apply
CPT symmetry	does not apply	applies	does not apply
renormalization	does not apply	applies	does not apply
permutation symmetry	does not apply	applies	does not apply
interactions	do not exist	exist	do not exist
number of particles	undefined	defined	undefined
algebras of observables	undefined	defined	undefined
matter indistinguishable from vacuum	true	false	true
boundaries exist	false	true	false
nature is a set	false	true	false

relating distances r or coordinates x_μ with their inverse values using the Planck length $l_{\rm Pl}$. Can this mapping, called *inversion*, be a symmetry of nature? At every point of space? For example, if the horizon distance is inserted, equation (689) implies that lengths smaller than $l_{\rm Pl}/10^{61} \approx 10^{-96}$ m never appear in physics. Is this the case? What would inversion imply for the big bang? Challenge 1421 ny

Numerous fascinating questions are contained in the simple hypothesis of extremal identity. They lead to *two* main directions for investigation.

First, we have to search for some stronger arguments for the validity of extremal identity. We will discover a number of simple arguments, all showing that extremal identity is indeed a property of nature and producing many beautiful insights.

The other quest then follows. We need to find the correct version of equation (689). That oversimplified expression is neither sufficient nor correct. It is not sufficient because it does not explain *any* of the issues left open by general relativity and quantum theory. It only *relates* some of them, thus reducing their number, but it does not *solve* any of them. You may wish to check this for yourself. In other words, we need to find the precise description of quantum geometry and of elementary particles. Challenge 1422 ny

However, inversion is also simply wrong. Inversion is not the correct description of extremal identity because it does not realize a central result discovered above: it does not connect *states* and *intrinsic properties*. Inversion keeps them distinct. This means that inversion does not take *interactions* into account. And most open issues at this point of our mountain ascent are properties of interactions.

Bibliography

1027 Available at http://www.dse.nl/motionmountain/C10-QMGR.pdf. Those pages are a reworked version of the ones published in French as C. Schiller, Le vide diffère-t-il de la matière? in E. Gunzig & S. Diner, editeurs, *Le vide – Univers du tout et du rien – Des physiciens et des philosophes s'interrogent*, Les Éditions de l'Université de Bruxelles, 1998. See also C. Schiller, Does matter differ from vacuum? http://www.arxiv.org/abs/gr-qc/9610066. Cited on pages 952, 956, and 962.

1028 See the lucid discussion by G.F.R. Ellis & T. Rothman, Lost horizons, *American Journal of Physics* **61**, pp. 883–893, 1993. Cited on pages 955, 958, 959, and 960.

1029 H. Salecker & E.P. Wigner, Quantum limitations of the measurement of space-time distances, *Physical Review* **109**, pp. 571–577, 1958. See also the popular account of this paper by E.J. Zimmerman, The macroscopic nature of space-time, *American Journal of Physics* **30**, pp. 97–105, 1962. Cited on page 955.

1030 F. Karolyhazy, Gravitation and quantum mechanics of macroscopic objects, *Il Nuovo Cimento* **A42**, pp. 390–402, 1966, Y.J. Ng & H. Van Dam, Limit to space-time measurement, *Modern Physics Letters A* **9**, pp. 335–340, 1994, Y.J. Ng & H. Van Dam, Remarks on gravitational sources, *Modern Physics Letters A* **10**, pp. 2801–2808, 1995. Newer summaries are Y.J. Ng & H. van Dam, Comment on "Uncertainty in measurements of distance", http://www.arxiv.org/abs/gr-qc/0209021 and Y.J. Ng, Spacetime foam, http://www.arxiv.org/abs/gr-qc/0201022. Cited on page 955.

1031 See, for example, the Hollywood movie Contact, based on the book by Carl Sagan, *Contact*, 1985. Cited on page 960.

1032 See, for example, the international bestseller by S. Hawking, *A Brief History of Time – From the Big Bang to Black Holes*, 1988. Cited on page 962.

1033 L. Rosenfeld, in H.-J. Treder, *Entstehung, Entwicklung und Perspektiven der Einsteinschen Gravitationstheorie*, Springer Verlag, 1966. Cited on page 965.

1034 Holography in high-energy physics is connected with the work by 't Hooft and by Susskind. See for example G. 't Hooft, Dimensional reduction in quantum gravity, pp. 284–296, in A. Ali, J. Ellis & S. Randjbar-Daemi, *Salaamfeest*, 1993, or the often cited paper by L. Susskind, The world as a hologram, *Journal of Mathematical Physics* **36**, pp. 6377–6396, 1995, or as http://www.arxiv.org/abs/hep-th/9409089 A good modern overview is R. Bousso, The holographic principle, *Review of Modern Physics* **74**, pp. 825–874, 2002, also available as http://www.arxiv.org/abs/hep-th/0203101, even though it has some argumentation errors, as explained on page 812. Cited on page 966.

1035 S. Lloyd, Computational capacity of the universe, *Physical Review Letters* **88**, p. 237901, 2002. Cited on page 968.

1036 Gottfried Wilhelm Leibniz, *La Monadologie*, 1714. Written in French, it is available freely at http://www.uqac.uquebec.ca/zone30/Classiques_des_sciences_sociales and in various other languages on other websites. Cited on page 968.

1037 See, for example, H. Wussing & P.S. Alexandrov, (eds.), *Die Hilbertschen Probleme*, Akademische Verlagsgesellschaft Geest & Portig, 1983, or Ben H. Yandell, *The Honours Class: Hilbert's Problems and their Solvers*, A.K. Peters, 2002. Cited on page 968.

1038 Large part of the study of dualities in string and M theory can be seen as investigations into the detailed consequences of extremal identity. For a review of duality, see ... A classical version of duality is discussed by M.C.B. Abdalla, A.L. Gadelka & I.V. Vancea, Duality between coordinates and the Dirac field, http://www.arxiv.org/abs/hep-th/0002217. Cited on page 971.

Sex is the physics urge sublimated.

Graffito

Maybe you have once met a physicist who has told you, in one of those moments of confidentiality, that studying physics is more beautiful than making love. At this statement, many will simply shake their head in disbelief and strongly disapprove. In this section we shall argue that it is possible to learn so much about physics while making love that discussions about their relative beauty can be put aside altogether.

Imagine to be with your partner on a beautiful tropical island, just after sunset, and to look together at the evening sky. Imagine as well that you know little of what is taught at school nowadays, e.g. that your knowledge is that of the late Renaissance, which probably is a good description of the average modern education level anyway.

Imagine being busy enjoying each other's company. The most important results of physics can be deduced from the following experimental facts:*

Love is communication.	Love is tiring.
Love is an interaction between moving bodies.	Love takes time.
Love is attractive.	Love is repulsive.
Love makes noise.	In love, size matters.
Love is for reproduction.	Love can hurt.
Love needs memory.	Love is Greek.
Love uses the sense of sight.	Love is animalic.
Love is motion.	Love is holy.
Love is based on touch.	Love uses motion again.
Love is fun.	Love is private.
Love makes one dream.	

Let us see what these observations imply for the description of nature.

▪ Love is *communication*. Communication is possible because nature looks similar from different standpoints and because nature shows no surprises. Without similarity we could not understand each other, and a world of surprises would even make thinking impossible; it would not be possible to form concepts to describe observations. But fortunately, the world is regular; it thus allows to use concepts such as time and space for its description.

▪ Love is an *interaction between moving bodies*. Together with the previous result, this implies that we can and need to describe moving bodies with *mass*, *energy* and *momentum*. That is not a small feat. For example, it implies that the Sun will rise tomorrow
Challenge 1423 n
if the sea level around the island is the usual one.

▪ Love is *attractive*. When feeling attracted to your partner, you may wonder if this attraction is the same which keeps the Moon going around the Earth. You make a quick

* Here we deduce physics from love. We could also deduce physics from sexuality. The modern habit of saying 'sex' instead of 'sexuality' mixes up two completely different concepts. In fact, studying the influences
Ref. 1039
of *sex* on physics is almost fully a waste of time. We avoid it. Maybe one day we shall understand why there do not seem to be any female crackpots proposing pet physical theories.

calculation and find that applying the expression for universal gravity

$$E_{\text{pot}} = -\frac{GMm}{r} \tag{690}$$

to both of you, the involved energy is about as much as the energy added by the leg of a fly on the skin. (M and m are your masses, r your distance, and the gravitational constant has the value $G = 6.7 \cdot 10^{-11}\,\mathrm{m^3/kg\,s^2}$.) In short, your partner teaches you that in nature there are other attractive interactions apart from gravity; the average modern education is incomplete.

Nevertheless, this first equation is important: it allows to predict the position of the planets, the time of the tides, the time of eclipses, the return of comets, etc., to a high accuracy for thousands of years in advance.

- Love *makes noise*. That is no news. However, even after making love, even when everybody and everything is quiet, in a completely silent environment, we do hear something. The noises we hear are produced within the ear, partly by the blood flowing through the head, partly by the electrical noise generated in the nerves. That is strange. If matter were continuous, there would be no noise even for low signal levels. In fact, all proofs for the discreteness of matter, of electric current, of energy, or of light are based on the increase of fluctuations with the smallness of systems under consideration. The persistence of noise thus makes us suspect that matter is made of smallest entities. Making love confirms this suspicion in several ways.
- Love is for *reproduction*. Love is what we owe our life to, as we all are results of reproduction. But the reproduction of a structure is possible only if it can be constructed, in other words if the structure can be built from small standard entities. Thus we again suspect ourselves to be made of smallest, discrete entities.

Love is also a complicated method of reproduction. Mathematics provides a much simpler one. If matter objects were not made of particles, but were continuous, it would be possible to perform reproduction by cutting and reassembling. A famous mathematical theorem by Banach and Tarski proves that it is possible to take a continuous solid, cut it into five pieces and rearrange the pieces in such a way that the result are *two* copies of the same size and volume as the original. In fact, even volume increases can be produced in this way, thus realizing growth without any need for food. Mathematics thus provides some interesting methods for growth and reproduction. However, they assume that matter is continuous, without a smallest length scale. The observation that these methods do not work in nature is compatible with the idea that matter is not continuous.

Ref. 1040

- Love *needs memory*. If you would not recognize your partner among all possible ones, your love life would be quite complicated. A memory is a device which, in order to store information, must have small internal fluctuations. Obviously, fluctuations in systems get smaller as their number of components increase. Since our memory works so well, we can follow that we are made of a large number of small particles.

In summary, love shows that we are made of some kind of lego bricks: depending on the level of magnification, these bricks are called molecules, atoms, or elementary particles. It is possible to estimate their size using the sea around the tropical island, as well as a bit of oil. Can you imagine how?

Challenge 1424 n

- Love *uses the sense of sight*. Seeing each other is only possible because we are cold whereas the Sun is hot. If we and our environment all had the same temperature as the Sun, we would not see each other. This can be checked experimentally by looking into a hot oven: Inside a glowing oven filled with glowing objects it is impossible to discern them against the background. Ref. 1041

▪ Love *is motion*. Bodies move against each other. Moreover, their speed can be measured. Since measurement is a comparison with a standard, there must be a velocity standard in nature, some special velocity standing out. Such a standard must either be the minimum or the maximum possible value. Now, daily life shows that for velocity a finite minimum value does not exist. We are thus looking for a maximum value. To estimate the value of the maximum, just take your mobile phone and ring home from the island to your family. From the delay in the line and the height of the satellite, you can deduce the telephone speed c and get $3 \cdot 10^8$ m/s.

The existence of a maximum speed c implies that time is different for different observers. Looking into the details, we find that this effect becomes noticeable at energies

$$E_{\text{different time}} \approx mc^2 , \tag{691}$$

where m is the mass of the object. For example, this applies to electrons inside a television tube.

▪ Love is based on *touching*. When we touch our partner, sometimes we get small shocks. The energies involved are larger than than those of touching fly legs. In short, people are electric.

In the dark, we observe that discharges emit light. Light is thus related to electricity. In addition, touching proves that light is a wave: simply observe the dark lines between two fingers near your eye in front of a bright background. The lines are due to interference effects. Light thus does not move with infinite speed. In fact, it moves with the same speed as that of telephone calls. Challenge 1425 n

▪ Love is *fun*. People like to make love in different ways, such as in a dark room. But rooms get dark when the light is switched off only because we live in a space of odd dimensions. In even dimensions, a lamp would not turn off directly after the switch is flipped, but dim only slowly. Page 193

Love is also fun because with our legs, arms and bodies we can make knots. Knots are possible only in three dimensions. In short, love is real fun only because we live in 3 dimensions.

▪ Love is *tiring*. The reason is gravity. But what is gravity? A little thinking shows that since there is a maximum speed, gravity is the curvature of space-time. Curved space also means that a *horizon* can appear, i.e. a largest possible visible distance. From equations (690) and (691), we deduce that this happens when distances are of the order of

$$R_{\text{horizon}} \approx Gm/c^2 . \tag{692}$$

For example, only due a horizon, albeit one appearing in a different way, the night sky is dark.

▪ Love *takes time*. It is known that men and women have different opinions on durations. It is also known that love happens between your ears. Indeed, biological research has shown that we have a clock inside the brain, due to circulating electrical currents. Ref. 1045

This clock provides our normal sense of time. Since such a brain clock can be built, there must be a time standard in nature. Again, such a standard must be either a minimum or a maximum time interval. We shall discover it later on.

▪ Love is *repulsive*. And in love, *size matters*. Both facts turn out to be the two sides of the same coin. Love is based on touch, and touch needs repulsion. Repulsion needs a length scale, but neither gravity nor classical electrodynamics provide one. Classical physics only allows for the measurement of speed. Classical physics cannot explain that the measurement of length, time, or mass is possible.* Classically, matter cannot be hard; it should be possible to compress it. But love shows us that this is not the case. Love shows us that lengths scales do exist in nature and thus that classical physics is not sufficient for the description of nature.

▪ Love can *hurt*. For example, it can lead to injuries. Atoms can get ripped apart. That happens when energies are concentrated on small volumes, such as a few aJ per atom. Investigating such situations more precisely, we finds that strange phenomena appear at distances r if energies exceed the value

$$E \approx \frac{\hbar\, c}{r} \quad ; \tag{693}$$

in particular, energy becomes chunky, things become fuzzy, boxes are not tight, and particles get confused. These are called *quantum* phenomena. The new constant $\hbar = 10^{-34}$ Js is important: it determines the size of things, because it allows to define distance and time units. In other words, objects tear and break because in nature there is a minimum action, given roughly by $\hbar$.

If even more energy is concentrated in small volumes, such as energies of the order of mc^2 per particle, one even observes transformation of energy into matter, or *pair production*. From equations (691) and (693), we deduce that this happens at distances of

$$r_{\text{pair production}} \approx \frac{\hbar}{m\, c} \ . \tag{694}$$

At such small distances we cannot avoid using the quantum description of nature.

▪ Love is not only *Greek*. The Greeks were the first to make theories above love, such as Plato in his *Phaedrus*. But they also described it in another way. Already before Plato, Democritus said that making love is an example of particles moving and interacting in vacuum. If we change 'vacuum' to 'curved 3+1-dimensional space' and 'particle' to 'quantum particle', we do indeed make love in the way Democritus described 2500 years ago.

It seems that physics has not made much progress in the mean time. Take the statement made in 1939 by the British astrophysicist Arthur Eddington:

> I believe there are 15,747,724,136,275,002,577,605,653,961,181,555,468,044,717,914-,527,116,709,366,231,425,076,185,631,031,296 protons in the universe and the same number of electrons.

Ref. 1042

* Note that the classical electron radius is not an exception: it contains the elementary charge e, which contains a length scale, as shown on page 455.

Compare it with the version of 2005:

> Baryons in the universe: $10^{81\pm1}$; total charge: near zero.

The second is more honest, but which of the two is less sensible? Both sentences show that there are unexplained facts in the Greek description nature, in particular the number of involved particles.

▪ Love is *animalic*. We have seen that we can learn a lot about nature from the existence of love. We could be tempted to see this approach of nature as a special case of the so-called *anthropic principle*. However, some care is required here. In fact, we could have learned exactly the same if we had taken as starting point the observation that *apes* or *pigs* have love. There is no 'law' of nature which distinguishes between them and humans. In fact, there is a simple way to determine whether any 'anthropic' statement makes sense: the reasoning must be equally true for humans, apes, and pigs.

A famous anthropic deduction was drawn by the British astrophysicist Fred Hoyle. While studying stars, he predicted a resonance in the carbon-12 nucleus. If it did not exist, he argued, stars could not have produced the carbon which afterwards was spread out by explosions into interstellar space and collected on Earth. Also apes or pigs could reason this way; therefore Hoyle's statement does make sense.

On the other hand, claiming that the universe is made *especially* for people is not sensible: using the same arguments, pigs would say it is made for pigs. The existence of either requires all 'laws' of nature. In summary, the anthropic principle is true only in so far as its consequences are indistinguishable from the porcine or the simian principle. In short, the animalic side of love puts limits to the philosophy of physics.

▪ Love is *holy*. Following the famous definition by the theologian Rudolf Otto, holiness
Ref. 1043 results from a mixture of a *mysterium tremendum* and a *mysterium fascinans*. Tremendum means that it makes one tremble. Indeed, love produces heat and is a dissipative process. All systems in nature which produce heat have a finite lifetime. That is true for machines, stars, animals, lightning, fire, lamps and people. Through heat, love shows us that we are going to die. Physicists call this the second principle of thermodynamics.

But love also fascinates. Everything which fascinates has a story. Indeed, this is a principle of nature: every dissipative structure, every structure which appears or is sustained through the release of energy, tells us that it has a story. Take atoms, for example. All the protons we are made of formed during the big bang. Most hydrogen we are made of is also that old. The other elements were formed in stars and then blown into the sky during nova or supernova explosions. They then regrouped during planet formation. We truly are made of stardust.

Why do such stories fascinate? If you only think about how you and your partner have met, you will discover that it is through a chain of incredible coincidences. If only one of all these coincidences had not taken place, you and your partner would not be together. And of course, we all owe our existence to such a chain of coincidences, which brought our parents together, our grandparents, and made life appear on Earth.

The realization of the importance of coincidences automatically produces two kinds of questions: *why?* and *what if?* Physicists have now produced a list of all the answers to repeated why questions and many are working at the list of what-if questions. The first list, the why-list of Table 73, gives all facts still unexplained. It can also be called the complete list of all surprises in nature. (Above, it was said that there are no surprises in nature about *what* happens. However, so far there still are a handful of surprises on *how* all these things happen.)

Table 73 *Everything* quantum field theory and general relativity do *not* explain; in other words, a list of *the only* experimental data and criteria available for tests of the unified description of motion

Observable	Property unexplained so far
Local quantities, from quantum theory	
α_{em}	the low energy value of the electromagnetic coupling constant
α_w	the low energy value of the weak coupling constant
α_s	the low energy value of the strong coupling constant
m_q	the values of the 6 quark masses
m_l	the values of 3 lepton masses
m_W	the values of the independent mass of the W vector boson
θ_W	the value of the Weinberg angle
$\beta_1, \beta_2, \beta_3$	three mixing angles
θ_{CP}	the value of the CP parameter
θ_{st}	the value of the strong topological angle
3	the number of particle generations
$0.5\,\mathrm{nJ/m^3}$	the value of the observed vacuum energy density or cosmological constant
$3+1$	the number of space and time dimensions
Global quantities, from general relativity	
$1.2(1) \cdot 10^{26}$ m ?	the distance of the horizon, i.e. the 'size' of the universe (if it makes sense)
10^{82} ?	the number of baryons in the universe, i.e. the average matter density in the universe (if it makes sense)
$> 10^{92}$?	the initial conditions for more than 10^{92} particle fields in the universe, including those at the origin of galaxies and stars (if they make sense)
Local structures, from quantum theory	
$S(n)$	the origin of particle identity, i.e. of permutation symmetry
Ren. group	the renormalization properties, i.e. the existence of point particles
SO(3,1)	the origin of Lorentz (or Poincaré) symmetry (i.e. of spin, position, energy, momentum)
C^*	the origin of the algebra of observables
Gauge group	the origin of gauge symmetry (and thus of charge, strangeness, beauty, etc.)
in particular, for the standard model:	
U(1)	the origin of the electromagnetic gauge group (i.e. of the quantization of electric charge, as well as the vanishing of magnetic charge)
SU(2)	the origin of weak interaction gauge group
SU(3)	the origin of strong interaction gauge group
Global structures, from general relativity	

Observable	Property unexplained so far
maybe $R \times S^3$?	the unknown topology of the universe (if it makes sense)

This why-list fascinates through its shortness, which many researchers are still trying to reduce. But it is equally interesting to study what consequences appear if any of the values from Table 73 were only a tiny bit different. It is not a secret that small changes in nature would lead to completely different observations, as shown in Table 74.

Table 74 A small selection of the consequences when changing aspects of nature

Observable	Change	Result
Moon size	smaller	small Earth magnetic field; too much cosmic radiation; widespread child cancers.
Moon size	larger	large Earth magnetic field; too little cosmic radiation; no evolution into humans.
Jupiter	smaller	too many comet impacts on Earth; extinction of animal life.
Jupiter	larger	too little comet impacts on Earth; no Moon; no dinosaur extinction.
Oort belt	smaller	no comets, no irregular asteroids, no Moon; still dinosaurs.
Star distance	smaller	irregular planet motion; supernova dangers.
Strong coupling constant	smaller	proton decay; leucemia.

The large number of coincidences of life force our mind to realize that we are only a *tiny* part of nature. We are a small droplet shaken around in the ocean of nature. Even the tiniest changes in nature would prevent the existence of humans, apes and pigs. In other words, making love tells us that the universe is much larger than we are and tells us how much we are dependent and connected to the rest of the universe.

▪ We said above that love uses *motion*. It contains a remarkable mystery, worth a second look:

- Motion is the change of position with time of some bodies.

- Position is what we measure with a ruler. Time is what we measure with a clock. Both rulers and clocks are bodies.

- A body is an entity distinct from its environment by its shape or its mass. Shape is the extension of a body in space (and time). Mass is measured by measuring speed or acceleration, i.e. by measuring space and time.

This means that we define space-time with bodies – as done in detail in general relativity – and that we define bodies with space-time – as done in detail in quantum theory. This circular reasoning shows that making love is truly a mystery. The circular reasoning has not yet been eliminated yet; at present, modern theoretical physicists are busy attempting to do so. The most promising approach seems to be M-theory, the modern extension of string theory. But any such attempt has to overcome important difficulties which can also be experienced while making love.

▪ Love is *private*. But is it? Privacy assumes that a person can separate itself from the rest, without important interactions, at least for a given time, and come back later. This is possible if the person puts enough *empty space* between itself and others. In other words, privacy is based on the idea that objects can be distinguished from vacuum. Let us check whether this is always possible.

Ref. 1044 What is the smallest measurable distance? This question has been almost, but only almost answered by Max Planck in 1899. The distance δl between two objects of mass m is surely larger than their position indeterminacy $\hbar/\Delta p$; and the momentum indeterminacy must be smaller that the momentum leading to pair production, i.e. $\Delta p < mc$. This means that

$$\delta l \geqslant \Delta l \geqslant \frac{\hbar}{mc} . \tag{695}$$

In addition, the measurements require that signals leave the objects; the two masses must not be black holes. Their masses must be so small that the Schwarzschild radius is smaller than the distance to be measured. This means that $r_S \approx Gm/c^2 < \delta l$ or that

$$\delta l \geqslant \sqrt{\frac{\hbar G}{c^3}} = l_{Pl} = 1.6 \cdot 10^{-35}\,\mathrm{m} . \tag{696}$$

This expression defines a minimum length in nature, the so-called *Planck length*. Every other Gedanken experiment leads to this characteristic length as well. In fact, this minimum distance (and the corresponding minimum time interval) provides the measurement standard we were looking for at the beginning of our musings about length and time measurements.

A more detailed discussion shows that the smallest *measurable* distance is somewhat larger, a multiple of the Planck length, as measurements require the distinction of matter and radiation. This happens at scales about 800 times the Planck length.

In other words, privacy has its limits. In fact, the issue is even more muddled when we explore the consequences for bodies. A body, also a human one, is something we can touch, throw, hit, carry or weigh. Physicists say that a body is something with energy or momentum. Vacuum has none of it. In addition, vacuum is unbounded, whereas objects are bounded.

What happens if we try to weigh objects at Planck scales? Quantum theory makes a simple prediction. If we put an object of mass M in a box of size R onto a scale – as in Figure 371 – equation (693) implies that there is a minimal mass error

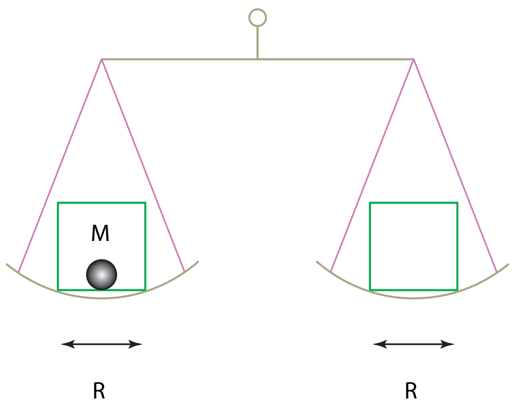


Figure 371 A *Gedanken experiment* showing that at Planck scales, matter and vacuum cannot be distinguished

ΔM given by

$$\Delta M \approx \frac{\hbar}{cR} . \qquad (697)$$

If the box has Planck size, the mass error is the Planck mass

$$\Delta M = M_{\rm Pl} = \sqrt{\hbar c/G} \approx 22\,\mu\text{g} . \qquad (698)$$

How large is the mass we can put into a box of Planck size? Obviously it is given by the maximum possible mass density. To determine it, imagine a planet and put a satellite in orbit around it, just skimming its surface. The density ρ of the planet with radius r is given by

$$\rho \approx \frac{M}{r^3} = \frac{v^2}{Gr^2} . \qquad (699)$$

Using equation (695) we find that the maximum mass density in nature, within a factor of order one, is the so-called *Planck density*, given by

$$\rho_{\rm Pl} = \frac{c^5}{G^2\hbar} = 5.2 \cdot 10^{96}\,\text{kg/m}^3 . \qquad (700)$$

Therefore the *maximum* mass that can be contained inside a Planck box is the Planck mass. But that was also the measurement *error* for that situation. This implies that we cannot say whether the original box we measured was empty or full: vacuum *cannot* be distinguished from matter at Planck scales. This astonishing result is confirmed by every other Gedanken experiment exploring the issue. Ref. 1044

It is straightforward to deduce with similar arguments that objects are not bound in size at Planck scales, i.e. that they are not localized, and that the vacuum is not necessarily extended at those scales. In addition, the concept of particle number cannot be defined at Planck scales. Challenge 1426 n

So, why is there something instead of nothing? Making love shows that there is no difference between the two options!

▪ Love makes us *dream*. When we dream, especially at night, we often look at the sky. How far is it away? How many atoms are enclosed by it? How old is it? These questions have an answer for small distances and for large distances; but for the whole of the sky or the whole of nature they cannot have one, as there is no way to be outside of the sky in order to measure it. In fact, each of the impossibilities to measure nature at smallest distances are found again at the largest scales. There seems to be a fundamental equivalence, or, as physicists say, a *duality* between the largest and the smallest distances. Ref. 1046

The coming years will hopefully show how we can translate these results into an even more precise description of motion and of nature. In particular, this description should allow us to reduce the number of unexplained properties of nature.

In summary, making love is a good physics lesson. Enjoy the rest of your day.

Bibliography

1039 An attempt to explain the lack of women in physics is made in Margaret Wertheim, *Pythagoras' Trousers – God, Physics and the Gender Wars*, Fourth Estate, 1997. Cited on page 975.

1040 The consequences of memory loss in this case are already told by Voltaire, *Aventure de la mémoire*, 1775. Cited on page 976.

1041 A picture of objects in a red hot oven and at room temperature is shown in C.H. Bennett, Demons, Engines and the Second Law, *Scientific American* **255**, pp. 108–117, November 1987. Cited on page 977.

1042 The famous quote is found at the beginning of chapter XI, 'The Physical Universe', in Arthur Eddington, *The Philosophy of Physical Science*, Cambridge, 1939. Cited on page 978.

1043 See the first intermezzo for details and references. Cited on page 979.

1044 Details can be found in Chapter XI of this text, a reworked version of the pages published in French as C. Schiller, Le vide diffère-t-il de la matière? in E. Gunzig & S. Diner, editeurs, *Le vide – Univers du tout et du rien – Des physiciens et des philosophes s'interrogent*, Les Éditions de l'Université de Bruxelles, 1998. Cited on pages 982 and 983.

1045 An introduction to the sense of time as result of signals in the ganglia and in the temporal lobe of the brain is found in R.B. Ivry & R. Spencer, The neural representation of time, *Current Opinion in Neurobiology* **14**, pp. 225–232, 2004. Cited on page 977.

1046 For details of fthe arguments leading to duality, see section 35. It also includes suggestions supporting the notion that the universe is not a even a set, thus proposing a solution for Hilbert's sixth problem. Cited on page 983.

37. Maximum force and minimum distance: physics in limit statements

The principle of maximum force or power allows us to summarize special relativity, quantum theory and general relativity in one fundamental limit principle each. The three principles are fully equivalent to the standard formulation of the theories. In particular, we show that the maximum force $c^4/4G$ implies and is equivalent to the field equations of general relativity. With the speed limit of special relativity and the action–angular momentum limit of quantum theory, the three fundamental principles imply a bound for every physical observable, from acceleration to size. The new, precise limit values differ from the usual Planck values by numerical prefactors of order unity.* Continuing this approach, we show that every observable in nature has both a lower and an upper limit value.

As a result, a maximum force and thus a minimum length imply that the non-continuity of space-time is an inevitable consequence of the unification of quantum theory and relativity. Furthermore, the limits are shown to imply the maximum entropy bound, including the correct numerical prefactor. The limits also imply a maximum measurement precision possible in nature, thus showing that any description by real numbers is approximate. Finally, the limits show that nature cannot be described by sets. As a result, the limits point to a solution to Hilbert's sixth problem. The limits also show that vacuum and matter cannot be fully distinguished in detail; they are two limit cases of the same entity. These fascinating results provide the basis for any search for a unified theory of motion.

Fundamental limits to all observables

At dinner parties physicists are regularly asked to summarize physics in a few sentences. It is useful to be able to present a few simple statements answering the request.** Here we propose such a set of statements. We are already familiar with each statement from what we have found out so far in our exploration of motion. But by putting them all together we will get direct insight into the results of modern research on unification.

The main lesson of modern physics is the following: Nature limits the possibilities of motion. These limits are the origin of special relativity, general relativity and quantum theory. In fact, we will find out that nature poses limits to *every* aspect of motion; but let us put first things first.

Special relativity in one statement

The step from everyday or Galilean physics to special relativity can be summarized in a single limit statement on motion. It was popularized by Hendrik Antoon Lorentz: *There is a maximum energy speed in nature.* For all physical systems and all observers,

$$v \leqslant c \,. \qquad (701)$$

* These limits were given for the first time in 2003, in this section of the present text.

** Stimulating discussions with Saverio Pascazio, Corrado Massa, Steve Carlip and Norbert Dragon helped shaping this section.

A few well-known remarks set the framework for the discussion that follows. The speed v is smaller than or equal to the speed of light c for *all* physical systems;* in particular, this limit is valid both for composed systems as well as for elementary particles. The speed limit statement is also valid for all observers; no exception to the statement is known. Indeed, only a maximum speed ensures that cause and effect can be distinguished in nature, or that sequences of observations can be defined. The opposite statement, implying the existence of (real) tachyons, has been explored and tested in great detail; it leads to numerous conflicts with observations.

The maximum speed forces us to use the concept of *space-time* to describe nature. Maximum speed implies that space and time mix. The existence of a maximum speed in Page 255 nature also implies observer-dependent time and space coordinates, length contraction, time dilation, mass–energy equivalence and all other effects that characterize special relativity. Only the existence of a maximum speed leads to the principle of maximum aging that governs special relativity; only a maximum speed leads to the principle of least action at low speeds. In addition, only a finite speed limit makes it to define a *unit* of speed that is valid at all places and at all times. If a global speed limit did not exist, no natural measurement standard for speed, independent of all interactions, would exist in nature; speed would not then be a measurable quantity.

Special relativity also limits the size of systems, independently of whether they are composed or elementary. Indeed, the limit speed implies that acceleration a and size l cannot be increased independently without bounds, because the two ends of a system must not interpenetrate. The most important case concerns massive systems, for which we have

$$l \leqslant \frac{c^2}{a} . \tag{702}$$

This size limit is induced by the speed of light c; it is also valid for the *displacement* d of a system, if the acceleration measured by an external observer is used. Finally, the limit implies an 'indeterminacy' relation

$$\Delta l \; \Delta a \leqslant c^2 \tag{703}$$

for the length and acceleration indeterminacies. You might want to take a minute to deduce it from the time-frequency indeterminacy. All this is standard knowledge. Challenge 1427 n

Quantum theory in one statement

In the same way, the difference between Galilean physics and quantum theory can be summarized in a single statement on motion, due to Niels Bohr: *There is a minimum action in nature.* For all physical systems and all observers, Ref. 1047 Page 656

$$S \geqslant \frac{\hbar}{2} . \tag{704}$$

* A *physical system* is a region of space-time containing mass-energy, the location of which can be followed over time and which interacts incoherently with its environment. With this definition, images, geometrical points or incomplete entangled situations are excluded from the definition of system.

Half the Planck constant $\hbar$ is the smallest observable action or angular momentum. This statement is valid for both composite and elementary systems. The action limit is used less frequently than the speed limit. It starts from the usual definition of the action, $S = \int (T - U) \mathrm{d}t$, and states that between two observations performed at times t and $t + \Delta t$, even if the evolution of a system is not known, the action is at least $\hbar/2$. Physical action is defined to be the quantity that measures the amount of change in the state of a physical system. In other words, there is always a minimum change of state taking place between two observations of a system. In this way, the quantum of action expresses the well-known fundamental fuzziness of nature at a microscopic scale.

It can easily be checked that no observation results in a smaller value of action, irrespective of whether photons, electrons or macroscopic systems are observed. No exception to the statement is known. A minimum action has been observed for fermions, bosons, laser beams and matter systems and for any combination of these. The opposite statement, implying the existence of change that is arbitrary small, has been explored in detail; Einstein's long discussion with Bohr, for example, can be seen as a repeated attempt by Einstein to find experiments that make it possible to measure arbitrarily small changes in nature. In every case, Bohr found that this aim could not be achieved.

Ref. 1048 The minimum value of action can be used to deduce the indeterminacy relation, the tunnelling effect, entanglement, permutation symmetry, the appearance of probabilities in quantum theory, the information theory aspect of quantum theory and the existence of elementary particle reactions. The minimum value of action implies that, in quantum theory, the three concepts of state, measurement operation and measurement result need to be distinguished from each other; a so-called *Hilbert space* needs to be introduced. The minimal action is also part of the Einstein–Brillouin–Keller quantization. The details of Page 656 these connections can be found in the chapter on quantum theory.

Obviously, the existence of a quantum of action has been known right from the beginning of quantum theory. The quantum of action is at the basis of all descriptions of quantum theory, including the many-path formulation and the information-theoretic descriptions. The existence of a minimum quantum of action is completely equivalent to all standard developments of quantum theory.

We also note that only a finite action limit makes it possible to define a *unit* of action. If an action limit did not exist, no natural measurement standard for action would exist in nature; action would not then be a measurable quantity.

The action bound $S \leqslant pd \leqslant mcd$, together with the quantum of action, implies a limit on the displacement d of a system between any two observations:

$$d \geqslant \frac{\hbar}{2mc} . \tag{705}$$

In other words, (half) the (reduced) Compton wavelength of quantum theory is recovered as lower limit on the displacement of a system, whenever gravity plays no role. Since the quantum *displacement* limit applies in particular to an elementary system, the limit is also valid for the *size* of a *composite* system. However, the limit is *not* valid for the size of *elementary* particles.

The action limit of quantum theory also implies Heisenberg's well-known indeterm- Page 661 inacy relation for the displacement d and momentum p of systems:

$$\Delta d \, \Delta p \geqslant \frac{\hbar}{2} \, . \qquad (706)$$

This is valid for both massless and massive systems. All this is textbook knowledge, of course.

General relativity in one statement

Least known of all is the possibility of summarizing the step from Galilean physics to general relativity in a single statement on motion: *There is a maximum force or power in nature*. For all physical systems and all observers,

$$F \leqslant \frac{c^4}{4G} = 3.0 \cdot 10^{43}\,\mathrm{N} \quad \text{or} \quad P \leqslant \frac{c^5}{4G} = 9.1 \cdot 10^{51}\,W \, . \qquad (707)$$

The limit statements contain both the speed of light c and the constant of gravitation G; they thus indeed qualify as statements concerning relativistic gravitation. Like the previous limit statements, they are valid for *all* observers. This formulation of general relativity is not common; in fact, it seems that it was only discovered 80 years after the theory of general relativity had first been proposed.* A detailed discussion is given in the chapter on general relativity. Ref. 1049 Page 319

The value of the force limit is the energy of a Schwarzschild black hole divided by its diameter; here the 'diameter' is defined as the circumference divided by π. The power limit is realized when such a black hole is radiated away in the time that light takes to travel along a length corresponding to the diameter.

Force is change of momentum; power is the change of energy. Since momentum and energy are conserved, force or power can be pictured as the flow of momentum or energy through a given surface. The value of the maximum force is the mass-energy of a black hole divided by its diameter. It is also the surface gravity of a black hole times its mass. The force limit thus means that no physical system of a given mass can be concentrated in a region of space-time *smaller* than a (non-rotating) black hole of that mass. In fact, the mass-energy concentration limit can be easily transformed by simple algebra into the force limit; both are equivalent. Challenge 1428 e

It is easily checked that the maximum force is valid for all systems *observed* in nature, whether they are microscopic, macroscopic or astrophysical. Neither the 'gravitational force' (as long as it is operationally defined) nor the electromagnetic or the nuclear interactions are ever found to exceed this limit. Challenge 1429 e

The next aspect to check is whether a system can be *imagined* that does exceed the limit. An extensive discussion shows that this is impossible, if the proper size of observers or test masses is taken into account. Even for a moving observer, when the force value is increased by the (cube of the) relativistic dilation factor, or for an accelerating observer, when the observed acceleration is increased by the acceleration of the observer itself, the force limit must still hold. However, no situations allow the limit to be exceeded, because Page 319

* It might be that the present author was the first to point it out, in this very textbook. Also Gary Gibbons found this result independently.

for high accelerations a, horizons appear at distance a/c^2; since a mass m has a minimum diameter given by $l \geqslant 4Gm/c^2$, we are again limited by the maximum force.

The exploration of the force and power limits shows that they are achieved only at horizons; the limits are not reached in any other situation. The limits are valid for all observers and all interactions.

As an alternative to the maximum force and power limits, we can use as basic principle the statement: *There is a maximum mass change in nature.* For all systems and observers, one has

$$\frac{\mathrm{d}m}{\mathrm{d}t} \leqslant \frac{c^3}{4G} = 1.0 \cdot 10^{35}\mathrm{kg/s} \; . \tag{708}$$

Equivalently, we can state: *There is a maximum mass per length ratio in nature.* For all systems and observers, one has

$$\frac{\mathrm{d}m}{\mathrm{d}l} \leqslant \frac{c^2}{4G} = 3.4 \cdot 10^{26}\mathrm{kg/m} \; . \tag{709}$$

In detail, both the force and the power limits state that the flow of momentum or of energy *through any physical surface* – a term defined below – of any size, for any observer, in any coordinate system, never exceeds the limit values. Indeed, as a result to the lack of nearby black holes or horizons, neither limit value is exceeded in any physical system found so far. This is the case at everyday length scales, in the microscopic world and in astrophysical systems. In addition, even Gedanken experiments do not allow to the
Page 319
limits to be exceeded. However, the limits become evident only when in such Gedanken experiments the size of observers or of test masses is taken into account. If this is not
Page 319
done, apparent exceptions can be constructed; however, they are then unphysical.

Deducing general relativity*

In order to elevate the force or power limit to a principle of nature, we have to show that, in the same way that special relativity results from the maximum speed, general relativity results from the maximum force.

Ref. 1050
The maximum force and the maximum power are only realized at horizons. Horizons are regions of space-time where the curvature is so high that it limits the possibility of observation. The name 'horizon' is due to a certain analogy with the usual horizon of everyday life, which also limits the distance to which one can see. However, in general relativity horizons are surfaces, not lines. In fact, we can *define* the concept of horizon in general relativity as a region of maximum force; it is then easy to prove that it is always a two-dimensional surface, and that it has all properties usually associated with it.

The connection between horizons and the maximum force or power allows a simple deduction of the field equations. We start with a simple connection. All horizons show energy flow at their location. This implies that horizons cannot be planes. An infinitely extended plane would imply an infinite energy flow. To characterize the finite extension of a given horizon, we use its radius R and its total area A.

* This section can be skipped at first reading.

The energy flow through a horizon is characterized by an energy E and a proper length L of the energy pulse. When such an energy pulse flows perpendicularly through a horizon, the momentum change $\mathrm{d}p/\mathrm{d}t = F$ is given by

$$F = \frac{E}{L} \; . \qquad (710)$$

For a horizon, we need to insert the maximum possible values. With the horizon area A and radius R, we can rewrite the limit case as

$$\frac{c^4}{4G} = \frac{E}{A} 4\pi R^2 \frac{1}{L} \qquad (711)$$

where the maximum force and the maximum possible area $4\pi R^2$ of a horizon of (maximum local) radius R were introduced. The fraction E/A is the energy per area flowing through the horizon. Often, horizons are characterized by the so-called surface gravity a instead of the radius R. In the limit case, two are related by $a = c^2/2R$. This leads to

$$E = \frac{1}{4\pi G} \, a^2 \, A \, L \; . \qquad (712)$$

Ref. 1051 Special relativity shows that horizons limit the product aL between proper length and acceleration to the value $c^2/2$. This leads to the central relation for the energy flow at horizons:

$$E = \frac{c^2}{8\pi G} \, a \, A \; . \qquad (713)$$

This equation makes three points. First, the energy flowing through a horizon is limited. Second, this energy is proportional to the area of the horizon. Third, the energy flow is proportional to the surface gravity. These results are fundamental statements of general relativity. No other part of physics makes comparable statements.

For the differential case the last relation can be rewritten as

$$\delta E = \frac{c^2}{8\pi G} \, a \, \delta A \; . \qquad (714)$$

In this way, the result can also be used for general horizons, such as horizons that are curved or time-dependent.*

Ref. 1052 In a well known paper, Jacobson has given a beautiful proof of a simple connection:

* Relation (714) is well known, though with different names for the observables. Since no communication is possible across a horizon, the detailed fate of energy flowing through a horizon is also unknown. Energy whose detailed fate is unknown is often called *heat*. Relation (714) therefore states that the heat flowing through a horizon is proportional to the horizon area. When quantum theory is introduced into the discussion, the area of a horizon can be called 'entropy' and its surface gravity can be called 'temperature'; relation (714) can then be rewritten as

$$\delta Q = T \delta S \; . \qquad (715)$$

However, this translation of relation (714), which requires the quantum of action, is unnecessary here. We if energy flow is proportional to horizon area for all observers and all horizons, then general relativity holds. To see the connection to general relativity, we generalize relation (714) to general coordinate systems and general energy-flow directions. This is achieved by introducing tensor notation.

To realize this at horizons, one introduces the general surface element $d\Sigma$ and the local boost Killing vector field k that generates the horizon (with suitable norm). Jacobson uses them to rewrite the left hand side of relation (714) as

$$\delta E = \int T_{ab} k^a d\Sigma^b \ , \tag{716}$$

where T_{ab} is the energy–momentum tensor. This rewrites the energy for arbitrary coordinate systems and arbitrary energy flow directions. Jacobson's main result is that the essential part of the right hand side of relation (714) can be rewritten, using the (purely geometric) Raychaudhuri equation, as

$$a\delta A = c^2 \int R_{ab} k^a d\Sigma^b \ , \tag{717}$$

where R_{ab} is the Ricci tensor describing space-time curvature.

Combining these two steps, we find that the energy–area relation (714) for horizons can be rewritten as

$$\int T_{ab} k^a d\Sigma^b = \frac{c^4}{8\pi G} \int R_{ab} k^a d\Sigma^b \ . \tag{718}$$

Jacobson shows that this equation, together with local conservation of energy (i.e., vanishing divergence of the energy–momentum tensor), can only be satisfied if

$$T_{ab} = \frac{c^4}{8\pi G} \left(R_{ab} - (\frac{1}{2} R + \Lambda) g_{ab} \right) \ , \tag{719}$$

where Λ is a constant of integration whose value is not specified by the problem. These are the full field equations of general relativity, including the cosmological constant Λ. The field equations are thus shown to be valid at horizons. Since it is possible, by choosing a suitable coordinate transformation, to position a horizon at any desired space-time event, the field equations must be valid over the whole of space-time.

It is possible to have a horizon at every event in space-time; therefore, at every event in nature there is the same maximum possible force (or power). This maximum force (or power) is a constant of nature.

In other words, we just showed that the field equations of general relativity are a direct consequence of the limited energy flow at horizons, which in turn is due to the existence of a maximum force or power. Maximum force or power implies the field equations. One can thus speak of the maximum force *principle*. In turn, the field equations imply maximum force. Maximum force and general relativity are equivalent.

The bounds on force and power have important consequences. In particular, they im-

only cite it to show the relation between horizon behaviour and quantum gravity.

ply statements on cosmic censorship, the Penrose inequality, the hoop conjecture, the non-existence of plane gravitational waves, the lack of space-time singularities, new experimental tests of the theory, and on the elimination of competing theories of relativistic
Ref. 1053 gravitation. These consequences are presented elsewhere.

Deducing universal gravitation

Universal gravitation can be derived from the force limit in the case where forces and speeds are much smaller than the maximum values. The first condition implies $\sqrt{4GMa} \ll c^2$, the second $v \ll c$ and $al \ll c^2$. Let us apply this to a specific case. We study a satellite circling a central mass M at distance R with acceleration a. This system, with length $l = 2R$, has only one characteristic speed. Whenever this speed v is much smaller than c, v^2 must be proportional both to $al = 2aR$ and to $\sqrt{4GMa}$. If they are taken together, they imply that $a = fGM/R^2$, where the numerical factor f is not yet
Challenge 1430 e fixed. A quick check, for example using the observed escape velocity values, shows that $f = 1$. Forces and speeds much smaller than the limit values thus imply that the inverse square law of gravity describes the interaction between systems. In other words, nature's limit on force implies the universal law of gravity, as is expected.

The size of physical systems in general relativity

General relativity, like the other theories of modern physics, provides a limit on the *size* of systems: there is a limit to the amount of matter that can be concentrated into a small volume.

$$l \geqslant \frac{4Gm}{c^2} \,. \tag{720}$$

The size limit is only achieved for black holes, those well-known systems which swallow everything that is thrown into them. It is fully equivalent to the force limit. All *composite* systems in nature comply with the lower size limit. Whether elementary particles fulfil or even achieve this limit remains one of the open issues of modern physics. At present, neither experiment nor theory allow clear statements on their size. More about this issue below.

Ref. 1054 General relativity also implies an 'indeterminacy relation':

$$\frac{\Delta E}{\Delta l} \leqslant \frac{c^4}{4G} \,. \tag{721}$$

Since experimental data are available only for composite systems, we cannot say yet whether this inequality also holds for elementary particles. The relation is not as popular as the previous. In fact, testing the relation, for example with binary pulsars, may lead to new tests that would distinguish general relativity from competing theories.

A mechanical analogy for the maximum force

The maximum force is central to the theory of general relativity. That is the reason why the value of the force (adorned with a factor 2π) appears in the field equations. The importance of a maximum force becomes clearer when we return to our old image of space-time as a deformable mattress. Like any material body, a mattress is characterized by a material constant that relates the deformation values to the values of applied energy. Similarly, a mattress, like any material, is characterized by the maximum stress it can bear before breaking. Like mattresses, crystals also have these two values. In fact, for perfect crystals (without dislocations) these two material constants are the same.

Empty space-time somehow behaves like a perfect crystal or a perfect mattress: it has a deformation-energy constant that at the same time is the maximum force that can be applied to it. The constant of gravitation thus determines the elasticity of space-time. Now, crystals are not homogeneous, but are made up of atoms, while mattresses are made up of foam bubbles. What is the corresponding structure of space-time? This is a central question in the rest of our adventure. One thing is sure: vacuum has no preferred directions, in complete contrast to crystals. In fact, all these analogies even suggest that the appearance of matter might be nature's way of preventing space-time from ripping apart. We have to patient for a while, before we can judge this option. A first step towards the answer to the question appears when we put all limits together.

Units and limit values for all physical observables

The existence of a maximum force in nature is equivalent to general relativity. As a result, physics can now be seen as making three simple statements on motion that is found in nature:

$$\begin{aligned} &\text{quantum theory on action:} && S \geqslant \frac{\hbar}{2} \\ &\text{special relativity on speed:} && v \leqslant c \\ &\text{general relativity on force:} && F \leqslant \frac{c^4}{4G} \;. \end{aligned} \tag{722}$$

The limits are valid for all physical systems, whether composed or elementary, and are valid for all observers. We note that the limit quantities of special relativity, quantum theory and general relativity can also be seen as the right-hand sides of the respective indeterminacy relations. Indeed, the set (703, 706, 721) of indeterminacy relations or the set (702, 705, 720) of length limits is each fully equivalent to the three limit statements (722). Each set of limits can be taken as a (somewhat selective) summary of twentieth century physics.

Challenge 1431 e

If the three fundamental limits are combined, a limit on a number of physical observables arise. The following limits are valid generally, for both composite and elementary systems:

$$\text{time interval:} \quad t \geqslant \sqrt{\frac{2G\hbar}{c^5}} = 7.6 \cdot 10^{-44}\,\mathrm{s} \tag{723}$$

$$\text{time distance product:} \quad td \geqslant \frac{2G\hbar}{c^4} = 1.7 \cdot 10^{-78}\,\mathrm{sm} \tag{724}$$

$$\text{acceleration:} \quad a \leqslant \sqrt{\frac{c^7}{2G\hbar}} = 4.0 \cdot 10^{51}\,\mathrm{m/s^2} \tag{725}$$

$$\text{angular frequency:} \qquad \omega \leqslant 2\pi\sqrt{\frac{c^5}{2G\hbar}} = 8.2 \cdot 10^{43}\,/\mathrm{s} \tag{726}$$

With the additional knowledge that, in nature, space and time can mix, we get

$$\text{distance:} \qquad d \geqslant \sqrt{\frac{2G\hbar}{c^3}} = 2.3 \cdot 10^{-35}\,\mathrm{m} \tag{727}$$

$$\text{area:} \qquad A \geqslant \frac{2G\hbar}{c^3} = 5.2 \cdot 10^{-70}\,\mathrm{m}^2 \tag{728}$$

$$\text{volume} \qquad V \geqslant \left(\frac{2G\hbar}{c^3}\right)^{3/2} = 1.2 \cdot 10^{-104}\,\mathrm{m}^3 \tag{729}$$

$$\text{curvature:} \qquad K \leqslant \frac{c^3}{2G\hbar} = 1.9 \cdot 10^{69}\,/\mathrm{m}^2 \tag{730}$$

$$\text{mass density:} \qquad \rho \leqslant \frac{c^5}{8G^2\hbar} = 6.5 \cdot 10^{95}\,\mathrm{kg/m}^3 \tag{731}$$

Of course, speed, action, angular momentum, power and force are also limited, as has already been stated. Except for a small numerical factor, for every physical observable these limits correspond to the Planck value. (The limit values are deduced from the commonly used Planck values simply by substituting $4G$ for G and $\hbar/2$ for $\hbar$.) These values are the true *natural units* of nature. In fact, the most aesthetically pleasing solution is to redefine the usual Planck values for every observable to these extremal values by absorbing the numerical factors into the respective definitions. In the following, we call the redefined limits the *(corrected) Planck limits* and assume that the factors have been properly included. In other words, *every natural unit or (corrected) Planck unit is at the same time the limit value of the corresponding physical observable.*

Most of these limit statements are found throughout the literature, though the numerical factors are often different. Each limit has a string of publications attached to it. The existence of a smallest measurable distance and time interval of the order of the Planck values is discussed in quantum gravity and string theory. The largest curvature has been studied in quantum gravity; it has important consequences for the 'beginning' of the universe, where it excludes any infinitely large or small observable. The maximum mass density appears regularly in discussions on the energy of vacuum.

Ref. 1055, Ref. 1056
Ref. 1057

With the present deduction of the limits, two results are achieved. First of all, the various arguments used in the literature are reduced to three generally accepted principles. Second, the confusion about the numerical factors is solved. During the history of Planck units, the numerical factors have varied greatly. For example, the fathers of quantum theory forgot the 1/2 in the definition of the quantum of action. Similarly, the specialists of relativity did not emphasize the factor 4. With the present framework, the issue of the correct factors in the Planck units can be considered as settled.

Limits to space and time

The three limits of nature (722) result in a minimum distance and a minimum time interval. These minimum intervals directly result from the unification of quantum theory and relativity. They do not appear if the theories are kept separate. In short, unification implies that there is a smallest length in nature. The result is important: *the formulation of physics as a set of limit statements shows that the continuum description of space and time is not correct.* Continuity and manifolds are only approximations; they are valid for large values of action, low speeds and low values of force. The reformulation of general relativity and quantum theory with limit statements makes this result especially clear. The result is thus a direct consequence of the unification of quantum theory and general relativity. No other assumptions are needed.

Ref. 1058 In fact, the connection between minimum length and gravity is not new. In 1969, Andrei Sakharov pointed out that a minimum length implies gravity. He showed that regularizing quantum field theory on curved space with a cut-off will induce counter-terms that include to lowest order the cosmological constant and then the Einstein Hilbert action.

The existence of limit values for the length observable (and all others) has numerous consequences discussed in detail elsewhere. (Ref. 1056, Ref. 1059) In particular, the existence of a smallest length – and a corresponding shortest time interval – implies that no surface is *physical* if any part of it requires a localization in space-time to dimensions smaller that the minimum length. (In addition, a physical surface must not cross any horizon.) Only by stipulation of this condition can unphysical examples that contravene the force and power limits be eliminated. For example, this condition has been overlooked in Bousso's earlier discussion of Bekenstein's entropy bound – though not in his more recent ones. (Ref. 1060)

The corrected value of the Planck length should also be the expression that appears in the so-called theories of 'doubly special relativity'. (Ref. 1061) These then try to expand special relativity in such a way that an invariant length appears in the theory.

A force limit in nature implies that no physical system can be smaller than a Schwarzschild black hole of the same mass. The force limit thus implies that *point particles do not exist.* So far, this prediction has not been contradicted by observations, as the predicted sizes are so small that they are outside experimental reach. If quantum theory is taken into account, this bound is sharpened. Because of the minimum length, elementary particles are now predicted to be larger than the corrected Planck length. Detecting the sizes of elementary particles would thus make it possible to check the force limit directly, for example with future electric dipole measurements. (Ref. 1056)

Mass and energy limits

Mass plays a special role in all these arguments. The set of limits (722) does not make it possible to extract a limit statement on the mass of physical systems. To find one, the aim has to be restricted.

The Planck limits mentioned so far apply for *all* physical systems, whether they are composed or elementary. Additional limits can only be found for elementary systems. In quantum theory, the distance limit is a size limit only for *composed* systems. A particle is elementary if the system size l is smaller than any conceivable dimension:

$$\text{for elementary particles:} \quad l \leqslant \frac{\hbar}{2mc} \,. \tag{732}$$

By using this new limit, valid only for elementary particles, the well-known mass, energy and momentum limits are found:

$$\begin{aligned}
&\text{for elementary particles:} \quad m \leqslant \sqrt{\frac{\hbar c}{8G}} &&= 7.7 \cdot 10^{-9}\,\mathrm{kg} = 0.42 \cdot 10^{19}\,\mathrm{GeV}/c^2 \\
&\text{for elementary particles:} \quad E \leqslant \sqrt{\frac{\hbar c^5}{8G}} &&= 6.9 \cdot 10^{8}\,\mathrm{J} = 0.42 \cdot 10^{19}\,\mathrm{GeV} \\
&\text{for elementary particles:} \quad p \leqslant \sqrt{\frac{\hbar c^3}{8G}} &&= 2.3\,\mathrm{kg\,m/s} = 0.42 \cdot 10^{19}\,\mathrm{GeV}/c
\end{aligned} \qquad (733)$$

These single-particle limits, corresponding to the corrected Planck mass, energy and momentum, were already discussed in 1968 by Andrei Sakharov, though again with different numerical factors. They are regularly cited in elementary particle theory. Obviously, all known measurements comply with the limits. It is also known that the unification of the electromagnetic and the two nuclear interactions takes place at an energy near, but still clearly below the maximum particle energy. Ref. 1062

Virtual particles – a new definition

In fact, there are physical systems that exceed all three limits. Nature does contain systems that move faster than light, that show action values below half the quantum of action and that experience forces larger than the force limit. The systems in question are called *virtual particles*.

Page 287 We know from special relativity that the virtual particles exchanged in collisions move faster than light. We know from quantum theory that virtual particle exchange implies action values below the minimum action. Page 709 Virtual particles also imply an instantaneous change of momentum; they thus exceed the force limit. *Virtual particles* are thus those particles that exceed all the limits that hold for (real) physical systems.

Limits in thermodynamics

Thermodynamics can also be summarized in a single statement on motion: *There is a smallest entropy in nature.*

$$S \geqslant \frac{k}{2} . \qquad (734)$$

The entropy S is limited by half the Boltzmann constant k. The result is almost 100 years old; it was stated most clearly by Leo Szilard, though with a different numerical factor. Ref. 1063 In the same way as in the other fields of physics, this result can also be phrased as a indeterminacy relation:

$$\Delta\frac{1}{T}\,\Delta U \geqslant \frac{k}{2} . \qquad (735)$$

Ref. 1064 This relation was given by Bohr and discussed by Heisenberg and many others (though with k instead of $k/2$). It is mentioned here in order to complete the list of indeterminacy relations and fundamental constants. With the single-particle limits, the entropy limit leads to an upper limit for temperature:

$$T \leqslant \sqrt{\frac{\hbar c^5}{2Gk^2}} = 1.0 \cdot 10^{32}\,\mathrm{K}\,. \tag{736}$$

This corresponds to the temperature at which the energy per degree of freedom is given by the (corrected) Planck energy. A more realistic value would have to take into account the number of degrees of freedom of a particle at Planck energy. This would change the numerical factor.

Electromagnetic limits and units

The discussion of limits can be extended to include electromagnetism. Using the (low-energy) electromagnetic coupling constant α, one gets the following limits for physical systems interacting electromagnetically:

electric charge: $$q \geqslant \sqrt{4\pi\varepsilon_0 \alpha c\hbar} = e = 0.16\,\mathrm{aC} \tag{737}$$

electric field: $$E \leqslant \sqrt{\frac{c^7}{64\pi\varepsilon_0\alpha\hbar G^2}} = \frac{c^4}{4Ge} = 2.4 \cdot 10^{61}\,\mathrm{V/m} \tag{738}$$

magnetic field: $$B \leqslant \sqrt{\frac{c^5}{64\pi\varepsilon_0\alpha\hbar G^2}} = \frac{c^3}{4Ge} = 7.9 \cdot 10^{52}\,\mathrm{T} \tag{739}$$

voltage: $$U \leqslant \sqrt{\frac{c^4}{32\pi\varepsilon_0\alpha G}} = \frac{1}{e}\sqrt{\frac{\hbar c^5}{8G}} = 1.5 \cdot 10^{27}\,\mathrm{V} \tag{740}$$

inductance: $$L \geqslant \frac{1}{8\pi\varepsilon_0\alpha}\sqrt{\frac{2\hbar G}{c^7}} = \frac{1}{e^2}\sqrt{\frac{\hbar^3 G}{2c^5}} = 4.4 \cdot 10^{-40}\,\mathrm{H} \tag{741}$$

With the additional assumption that in nature at most one particle can occupy one Planck volume, one gets

charge density: $$\rho_{\mathrm{e}} \leqslant \sqrt{\frac{\pi\varepsilon_0\alpha}{2G^3}}\,\frac{c^5}{\hbar} = e\sqrt{\frac{c^9}{8G^3\hbar^3}} = 1.3 \cdot 10^{85}\,\mathrm{C/m^3} \tag{742}$$

capacitance: $$C \geqslant 8\pi\varepsilon_0\alpha\sqrt{\frac{2\hbar G}{c^3}} = e^2\sqrt{\frac{8G}{c^5\hbar}} = 1.0 \cdot 10^{-46}\,\mathrm{F} \tag{743}$$

For the case of a single conduction channel, one gets

electric resistance: $$R \geqslant \frac{1}{8\pi\varepsilon_0\alpha c} = \frac{\hbar}{2e^2} = 2.1\,\mathrm{k\Omega} \tag{744}$$

electric conductivity: $$G \leqslant 8\pi\varepsilon_0\alpha c = \frac{2e^2}{\hbar} = 0.49\,\mathrm{mS} \tag{745}$$

electric current: $$I \leqslant \sqrt{\frac{2\pi\varepsilon_0\alpha c^6}{G}} = e\sqrt{\frac{c^5}{2\hbar G}} = 7.4 \cdot 10^{23}\,\mathrm{A} \tag{746}$$

Many of these limits have been studied already. The magnetic field limit plays a role in the discussion of extreme stars and black holes. The maximum electric field plays a role Ref. 1065 in the theory of gamma ray bursters. The studies of limit values for current, conductivity and resistance in single channels are well known; the values and their effects have been Ref. 1066 studied extensively in the 1980s and 1990s. They will probably win a Nobel prize in the future.

The observation of quarks and of collective excitations in semiconductors with charge $e/3$ does not necessarily invalidate the charge limit for physical systems. In neither case is there is a physical system – defined as localized mass-energy interacting incoherently with the environment – with charge $e/3$.

Vacuum and mass-energy – two sides of the same coin

In this discussion we have found that there is a fixed limit to every physical observable. Page 918 Many consequences have been discussed already in previous sections. There we saw that Page 952 the existence of a limit to all observables implies that at Planck scales no physical observable can be described by real numbers and that no low-energy symmetry is valid.

One conclusion is especially important for the rest of our adventure. We saw that there is a limit for the precision of length measurements in nature. The limit is valid both for the length measurements of empty space and for the length measurements of matter (or radiation). Now let us recall what we do when we measure the length of a table with a ruler. To find the ends of the table, we must be able to distinguish the table from the surrounding air. In more precise terms, we must be able to distinguish matter from vacuum. But we have no way to perform this distinction at Planck energy. In these domains, the intrinsic measurement limitations of nature do not allow us to say whether we are measuring vacuum or matter. There is no way, at Planck scales, to distinguish the two. Ref. 1056, Ref. 1059

Page 319 We have explored this conclusion in detail above and have shown that it is the only consistent description at Planck scales. The limitations in length measurement precision, in mass measurement precision and in the precision of any other observable do not allow to tell whether a box at Planck scale is full or empty. You can pick any observable you want to distinguish vacuum from matter. Use colour, mass, size, charge, speed, angular momentum, or anything you want. At Planck scales, the limits to observables also lead to limits in measurement precision. At Planck scales, the measurement limits are of the same size as the observable to be measured. As a result, it is impossible to distinguish matter and vacuum at Planck scales.

We put the conclusion in the sharpest terms possible: *Vacuum and matter do not differ at Planck scales.* This counter-intuitive result is one of the charms of theoretical high energy physics. This result alone inspired many researchers in the field and induced them to write best-sellers. Brian Greene was especially successful in presenting this side of Ref. 1067 quantum geometry to the wider public.

However, at this point of our adventure, most issues are still open. The precise manner in which a minimum distance leads to a homogeneous and isotropic vacuum is unclear. The way to describe matter and vacuum with the same concepts has to be found. And of course, the list of open questions in physics, given above, still waits. However, the conceptual results give hope; there are interesting issues awaiting us.

Paradoxes and curiosities about Planck limits

The (corrected) Planck limits are statements about properties of nature. There is no way to measure values exceeding these limits, whatever experiment is performed. As can be expected, such a claim provokes the search for counterexamples and leads to many paradoxes.

▪ The minimum angular momentum may surprise at first, especially when one thinks about particles with spin zero. However, the angular momentum of the statement is *total* angular momentum, including the orbital part with respect to the observer. The total angular momentum is never smaller than $\hbar/2$.

▪ If any interaction is stronger than gravity, how can the maximum force be determined by gravity alone, which is the weakest interaction? It turns out that in situations near the maximum force, the other interactions are negligible. This is the reason that gravity must be included in a unified description of nature.

▪ At first sight, it seems that electric charge can be used in such a way that the acceleration of a charged body towards a charged black hole is increased to a value exceeding the force limit. However, the changes in the horizon for charged black holes prevent this. Challenge 1432 ny

▪ The general connection that to every limit value in nature there is a corresponding indeterminacy relation is also valid for electricity. Indeed, there is an indeterminacy relation for capacitors of the form

$$\Delta C \; \Delta U \geqslant e \qquad (747)$$

where e is the positron charge, C capacity and U potential difference, and one between electric current I and time t

$$\Delta I \; \Delta t \geqslant e \qquad (748)$$

Ref. 1068 and both relations are found in the literature.

▪ The gravitational attraction between two masses never yields force values sufficiently high to exceed the force limit. Why? First of all, masses m and M cannot come closer than the sum of their horizon radii. Using $F = GmM/r^2$ with the distance r given by the (naive) sum of the two black hole radii as $r = 2G(M+m)/c^2$, one gets

$$F \leqslant \frac{c^4}{4G} \frac{Mm}{(M+m)^2} \,, \qquad (749)$$

which is never larger than the force limit. Even two attracting black holes thus do not exceed the force limit – in the inverse square approximation of universal gravity. The minimum size of masses does not allow to exceed the a maximum force.

▪ It is well known that gravity bends space. To be fully convincing, the calculation needs to be repeated taking space curvature into account. The simplest way is to study the force generated by a black hole on a test mass hanging from a wire that is lowered towards a black hole horizon. For an *unrealistic point mass*, the force would diverge on the horizon. Ref. 1069 Indeed, for a point mass m lowered towards a black hole of mass M at (conventionally defined radial) distance d, the force would be

$$F = \frac{GMm}{d^2\sqrt{1 - \frac{2GM}{dc^2}}} . \qquad (750)$$

The expression diverges at $d = 0$, the location of the horizon. However, even a test mass cannot be smaller than its own gravitational radius. If we want to reach the horizon with a *realistic* test mass, we need to chose a small test mass m; only a small – and thus light – mass can get near the horizon. For vanishingly small masses however, the resulting force tends to zero. Indeed, letting the distance tend to the smallest possible value by letting $d = 2G(m + M)/c^2 \to d = 2GM/c^2$ requires $m \to 0$, which makes the force $F(m, d)$ vanish. If on the other hand, we remain away from the horizon and look for the maximum force by using a mass as large as can possibly fit into the available distance (the calculation is straightforward algebra) again the force limit is never exceeded. In other words, for *realistic* test masses, expression (750) is *never* larger than $c^4/4G$. Taking into account the minimal size of test masses thus prevents that the maximum force is exceeded in gravitational systems.

▪ An absolute power limit implies a limit on the energy that can be transported per time unit through any imaginable surface. At first sight, it may seem that the combined power emitted by two radiation sources that each emit 3/4 of the maximum value should give 3/2 times the upper value. However, the combination forms a black hole or at least prevents part of the radiation to be emitted by swallowing some of it between the sources.

▪ One possible system that actually achieves the power limit is the final stage of black hole evaporation. But even in this case the power limit is not exceeded.

▪ The maximum force limit states that the stress–energy tensor, when integrated over any physical surface, does not exceed the limit value. No such integral, over any physical surface whatsoever, of any tensor component in any coordinate system, can exceed the force limit, provided that it is measured by a nearby observer or a test body with a realistic proper size. The maximum force limit thus applies to any component of any force vector, as well as to its magnitude. It applies to gravitational, electromagnetic, and nuclear forces. It applies to all realistic observers. Whether the forces are real or fictitious is not important. It also plays no role whether we discuss 3-forces of Galilean physics or 4-forces of special relativity. Indeed, the force limit applied to the 0-th component of the 4-force is the power limit.

Ref. 1049

▪ Translated to mass flows, the power limit implies that flow of water through a tube is limited in throughput. Also this limit seems unknown in the literature.

▪ The force limit cannot be overcome with Lorentz boosts. A Lorentz boost of any non-vanishing force value seems to allow exceeding the force limit at high speeds. However, such a transformation would create a horizon that makes any point with a potentially higher force value inaccessible.

▪ The power limits is of interest if applied to the universe as a whole. Indeed, it can be used to explain Olber's paradox. The sky is dark at night because the combined luminosity of all light sources in the universe cannot be brighter than the maximum value.

▪ One notes that the negative energy volume density $-\Lambda c^4/4\pi G$ introduced by the positive cosmological constant Λ corresponds to a negative pressure (both quantities have the same dimensions). When multiplied with the minimum area it yields a force value

$$F = \frac{\Lambda \hbar c}{2\pi} = 4.8 \cdot 10^{-79} N \ . \tag{751}$$

This is also the gravitational force between two corrected Planck masses located at the cosmological distance $\sqrt{\pi/4\Lambda}$. If we make the (wishful) assumption that this is the smallest possible force in nature (the numerical prefactor is not finalized yet), we get the fascinating conjecture that the full theory of general relativity, including the cosmological constant, is defined by the combination of a maximum and a minimum force in nature.

▪ Another consequence of the limits merits a dedicated section.

Upper and lower limits to observables

In our quest to understand motion, we have focused our attention to the limitations it is subjected to. Special relativity poses a limit to speed, namely the speed of light c. General Page 319 relativity limits force and power respectively by $c^4/4G$ and $c^5/4G$, and quantum theory introduces a smallest value $\hbar/2$ for angular momentum or action. We saw that nature poses a limit on entropy and a limit on electrical charge. Together, all these limits provide an extreme value for *every* physical observable. The extremum is given by the corresponding (corrected) Planck value. We have explored the list of all Planck limits in the previous Page 985 section. The question may now arise whether nature provides a limit for physical observables also on the opposite end of the measurement scale. For example, there is a highest force and a highest power in nature; is there also a lowest force and a lowest power? Is there also a lowest speed?

We show in the following that there indeed are such limits, for all observables. We give the general method to generate such bounds and explore several examples.* The exploration will lead us along an interesting scan across modern physics.

Size and energy dependence

Looking for additional limits in nature, we directly note a fundamental property. Any upper limit for angular momentum or any lower limit for power must be *system dependent*. Such limits will not be absolute, but will depend on properties of the system. Now, any physical system is a part of nature characterized by a boundary and its content.** The simplest properties all systems share are thus their size (characterized in the following by the diameter) L and energy E. With these characteristics we can enjoy deducing system-dependent limits for every physical observable. The general method is straightforward. We take the known inequalities for speed, action, power, charge and entropy and then extract a limit for any observable, by inserting length and energy as required. We then have to select the strictest of the limits we find.

* This section was added in June 2004.

** We mention here that quantum theory narrows down this definition as a part of nature that in addition Page 746 interacts *incoherently* with its environment. We assume that this condition is realized in the following.

Angular momentum, action and speed

It only takes a moment to note that the ratio of angular momentum D to mass times length has the dimension of speed. Since speeds are limited by the speed of light, we get

$$D \leqslant \frac{1}{c} LE \; . \tag{752}$$

Indeed, there do not seem to be any exceptions to this limit in nature. No known system has a larger angular momentum value, from atoms to molecules, from ice skaters to galaxies. For example, the most violently rotating objects, the so-called extremal black holes, are also limited in angular momentum by $D \leqslant LE/c$. (In fact, this limit is correct only if the energy is taken as the irreducible mass times c^2; if the usual mass is used, the limit is too large by a factor 4.) One remarks that the limit deduced from general relativity, given by $D \leqslant L^2c^3/4G$ is not stricter than the one just given. In addition, no system-dependent lower limit for angular momentum can be deduced.

Ref. 1070

The maximum angular momentum value is also interesting when it is seen as action limit. Action is the time integral of the difference between kinetic and potential energy. In fact, since nature always minimizes action W, we are not used to search for systems which maximize its value. You might check by yourself that the action limit $W \leqslant LE/c$ is not exceeded in any physical process.

Challenge 1433 ny

Similarly, speed times mass times length is an action. Since action values in nature are limited from below by $\hbar/2$, we get

$$v \geqslant \frac{\hbar c^2}{2} \frac{1}{LE} \; . \tag{753}$$

This relation is a rewritten form of the indeterminacy relation of quantum theory and is no news. No system of energy E and diameter L has a smaller speed than this limit. Even the slowest imaginable processes show this speed value. For example, the extremely slow radius change of a black hole by evaporation just realizes this minimal speed. Continuing with the method just outlined, one also finds that the limit deduced from general relativity, $v \leqslant (c^2/4G)(L/E)$, gives no new information. Therefore, no system-dependent upper speed limit exists.

Challenge 1434 ny

Challenge 1435 e

Incidentally, the limits are not unique. Additional limits can be found in a systematic way. Upper limits can be multiplied, for example, by factors of $(L/E)(c^4/4G)$ or $(LE)(2/\hbar c)$ yielding additional, but less strict upper limits. A similar rule can be given for lower limits.*

With the same approach we can now systematically deduce all size and energy dependent limits for physical observables. We have a tour of the most important ones.

* The strictest upper limits are thus those with the smallest possible exponent for length, and the strictest lower limits are those with the largest sensible exponent of length.

Force, power and luminosity

We saw that force and power are central to general relativity. Due to the connection $W = FLT$ between action W, force, distance and time, we can deduce

$$F \geqslant \frac{\hbar}{2c} \frac{1}{T^2} . \qquad (754)$$

Experiments do not reach this limit. The smallest forces measured in nature are those in atomic force microscopes, where values as small as 1 aN are observed. However, the values are all above the lower force limit.

The power P emitted by a system of size L and mass M is limited by

$$c^3 \frac{M}{L} \geqslant P \geqslant 2\hbar G \frac{M}{L^3} \qquad (755)$$

The left, upper limit gives the upper limit for any engine or lamp deduced from relativity; not even the universe exceeds it. The right, lower limit gives the minimum power emitted by any system due to quantum gravity effects. Indeed, no system is completely tight. Even black holes, the systems with the best ability in nature to keep components inside their enclosure, nevertheless radiate. The power radiated by black holes should just saturate this limit, provided the length L is taken to be the circumference of the black hole. (However, present literature values of the numerical factors in the black hole power are not yet consistent). The claim of the quantum gravity limit is thus that the power emitted by a black hole is the smallest power that is emitted by any composed system of the same surface gravity.

Acceleration

When the acceleration of a system is measured by a *nearby* observer, the acceleration a of a system of size L and mass M is limited by

$$c^2 \frac{1}{L} \geqslant a \geqslant 4G \frac{M}{L^2} \qquad (756)$$

The lower, right limit gives the acceleration due to universal gravity. Indeed, the relative acceleration between a system and an observer has at least this value. The left, upper limit to acceleration is the value due to special relativity. No exception to either of these limits has ever been observed. Using the limit to the size of masses, the upper limit can be transformed into the equivalent acceleration limit (Ref. 1071)

$$\frac{2c^3}{\hbar} M \geqslant a \qquad (757)$$

which has never been approached either, despite many attempts. The upper limit to acceleration is thus a quantum limit, the lower one a gravitational limit.

The acceleration of the radius of a black hole due to evaporation can be much slower

Challenge 1436 ny than the limit $a \geqslant 4GM/L^2$. Why is this not a counterexample?

Momentum

The momentum p of a system of size L is limited by

$$\frac{c^3}{4G} L \geqslant p \geqslant \frac{\hbar}{2} \frac{1}{L} \tag{758}$$

The lower limit is obviously due to quantum theory; experiments confirmed it for all radiation and matter. The upper limit for momentum is due to general relativity. It has never been exceeded.

Lifetime, distance and curvature

Time is something special. What are the limits to time measurements? Like before, we find that any measured time interval t in a system in thermal equilibrium is limited by

$$\frac{2}{\hbar} ML^2 \geqslant t \geqslant \frac{4G}{c^3} M \tag{759}$$

The lower time limit is the gravitational time. No clock can measure a smaller time than this value. Similarly, no system can produce signals shorter than this duration. Black holes, for example, emit radiation with a frequency given by this minimum value. The upper time limit is expected to be the exact lifetime of a black hole. (There is no consensus in the literature on the numerical factor yet.)

The upper limit to time measurements is due to quantum theory. It leads to a question: What happens to a single atom in space after the limit time has passed by? Obviously, an atom is not a composed system comparable with a black hole. The lifetime expression assumes that decay can take place in the most tiny energy steps. As long as there is no decay mechanism, the life-time formula does not apply. The expression (759) thus does not apply to atoms.

Distance limits are straightforward.

$$\frac{2c}{\hbar} ML^2 \geqslant d \geqslant \frac{4G}{c^2} M \tag{760}$$

Since curvature is an inverse square distance, curvature of space-time is also limited.

Mass change

The mass change $\mathrm{d}M/\mathrm{d}t$ of a system of size L and mass M is limited by

$$\frac{c^5}{16G^2} \frac{L}{M} \geqslant \frac{\mathrm{d}M}{\mathrm{d}t} \geqslant \frac{\hbar}{2} \frac{1}{L^2} \tag{761}$$

The limits apply to systems in thermal equilibrium. The left, upper limit is due to general relativity; it is never exceeded. The right, lower limit is due to quantum theory. Again, all experiments are consistent with the limit values.

Mass and density

Limits for rest mass make only sense if the system is characterized by a size L only. We then have

$$\frac{c^2}{4G} L \geqslant M \geqslant \frac{\hbar}{2c} \frac{1}{L} \tag{762}$$

The upper limit for mass was discussed in general relativity. Adding mass or energy to a black hole always increases its size. No system can show higher values than this value, and indeed, no such system is known or even imaginable. The lower limit on mass is obviously due to quantum theory; it follows directly from the quantum of action.

The mass density a system of size L is limited by

$$\frac{c^2}{4G} \frac{1}{L^2} \geqslant \rho \geqslant \frac{\hbar}{2c} \frac{1}{L^4} \tag{763}$$

The upper limit for mass density, due to general relativity, is only achieved for black holes. The lower limit is the smallest density of a system due to quantum theory. It also applies to the vacuum, if a piece of vacuum of site L is taken as a physical system. We note again that many equivalent (but less strict) limits can be formulated by using the transformations Page 1002 rules mentioned above.

The strange charm of the entropy bound

Ref. 1072 In 1973, Bekenstein discovered a famous limit that connects the entropy S of a physical system with its size and mass.

No system has a larger entropy than one bounded by a horizon. The larger the horizon surface, the larger the entropy. We write

$$\frac{S}{S_{\text{limit}}} \leqslant \frac{A}{A_{\text{limit}}} \tag{764}$$

which gives

$$S \leqslant \frac{kc^3}{4G\hbar} A , \tag{765}$$

where A is the surface of the system. Equality is realized only for black holes. The old question of the origin of the factor 4 in the entropy of black holes is thus answered here in he following way: it is due to the factor 4 in the force or power bound in nature (provided Ref. 1073 that the factors from the Planck entropy and the Planck action cancel). The future will Page 224 tell whether this explanation will stand the winds of time. Stay tuned.

We can also derive a more general relation if we use a mysterious assumption that we discuss afterwards. We assume that the limits for vacuum are opposite to those for matter.

We can then write $c^2/4G \leqslant M/L$ for the vacuum. This gives

$$S \leqslant \frac{\pi k c}{\hbar} ML = \frac{2\pi k c}{\hbar} MR \,. \tag{766}$$

In other words, we used

$$\frac{S}{S_{\text{corr. Planck}}} \leqslant \frac{M}{M_{\text{corr. Planck}}} \frac{A}{A_{\text{corr. Planck}}} \frac{L_{\text{corr. Planck}}}{L} \,. \tag{767}$$

Expression (766) is called *Bekenstein's entropy bound*. Up to today, no exception has been found or constructed, despite many attempts. Again, the limit value itself is only realized for black holes. Ref. 1060

We still need to explain the strange assumption used above. We are exploring the entropy of a horizon. Horizons are not matter, but limits to empty space. The entropy of horizons is due to the large amount of virtual particles found there. In order to deduce the maximum entropy of expression (767) one therefore has to use the properties of the vacuum. In other words, either (1) we use a mass to length ratio for vacuum *above* the Planck limit or (2) we use the Planck entropy as *maximum* value for vacuum.

Other, equivalent limits for entropy can be found if other variables are introduced. For example, since the ratio of shear viscosity η to the volume density of entropy (times k) has the dimensions of an action, we can directly write Ref. 1074

$$S \leqslant \frac{k}{\hbar} \eta V \tag{768}$$

Again, equality is only reached in the case of black holes. With time, the list of similar bounds will grow longer and longer.

Is there also a smallest entropy limit? So far, there does not seem to be a system-dependent minimum value for entropy; the approach gives no expression that is larger than k. Challenge 1437 ny

The entropy limit is an important step in making the description of motion consistent. If space-time can move, as general relativity maintains, it also has an entropy. How could entropy be limited if space-time is continuous? Clearly, due to the minimum distance and a minimum time in nature, space-time is not continuous, but has a finite number of degrees of freedom. The number of degrees of freedom and thus the entropy of space-time is thus finite.

In addition, the Bekenstein limit also allows some interesting speculations. Let us speculate that the universe itself, being surrounded by a horizon, saturates the Bekenstein bound. The entropy bound gives a bound to all degrees of freedom inside a system; it tells us that the number $N_{\text{d.o.f.}}$ of degrees of freedom of the universe is roughly Challenge 1438 e

$$N_{\text{d.o.f.}} \approx 10^{132} \,. \tag{769}$$

This compares with the number $N_{\text{Pl. vol.}}$ of Planck volumes in the universe

$$N_{\text{Pl. vol.}} \approx 10^{183} \tag{770}$$

and with the number $N_{\text{part.}}$ of particles in the universe

$$N_{\text{part.}} \approx 10^{91} . \tag{771}$$

In other words, particles are only a tiny fraction of what moves around. Most motion must be that of space-time. At the same time, space-time moves much less than naively expected. Finding out how all this happens is the challenge of the unified description of motion.

Temperature

A lower limit for the temperature of a thermal system can be found using the idea that the number of degrees of freedom of a system is limited by its surface, more precisely, by the ratio between the surface and the Planck surface. One gets the limit

$$T \geqslant \frac{4G\hbar}{\pi k c} \frac{M}{L^2} \tag{772}$$

Alternatively, using the method given above, one can use the limit on the thermal energy $kT/2 \geqslant \hbar c/(2\pi L)$ (the thermal wavelength must be smaller than the size of the system) together with the limit on mass $c^2/4G \geqslant M/L$ and deduce the same result.

We know the limit already: when the system is a black hole, it gives the temperature of the emitted radiation. In other words, the temperature of black holes is the lower limit for all physical systems for which a temperature can be defined, provided they share the *same boundary gravity*. As a criterion, boundary gravity makes sense: boundary gravity is accessible from the outside and describes the full physical system, since it makes use both of its boundary and its content. So far, no exception to this claim is known. All systems from everyday life comply with it, as do all stars. Even the coldest known systems in the universe, namely Bose–Einstein condensates and other cold matter in various laboratories, are much hotter than the limit, and thus much hotter than black holes of the same surface gravity. (Since a consistent Lorentz transformation for temperature is not possible, as we saw earlier, the limit of minimum temperature is only valid for an observer at the same gravitational potential and at zero relative speed to the system under consideration.)

Challenge 1439 n
Page 282

There seems to be no consistent way to define an upper limit for a system-dependent temperature. However, limits for other thermodynamic quantities can be found, but are not discussed here.

Challenge 1440 ny

Electromagnetic observables

When electromagnetism plays a role, the involved system also needs to be characterized by a charge Q. The method used so far then gives the following lower limit for the electric field E:

$$E \geqslant 4Ge \frac{M^2}{Q^2 L^2} \tag{773}$$

We write the limit using the elementary charge e, though writing it using the fine structure constant via $e = \sqrt{4\pi\varepsilon_0 \alpha \hbar c}$ would be more appropriate. Experimentally, this limit is not exceeded in any system in nature. Can you show whether it is achieved by maximally charged black holes?

Challenge 1441 ny

For the magnetic field we get

$$B \geqslant \frac{4Ge}{c} \frac{M^2}{Q^2 L^2} \tag{774}$$

Again, this limit is satisfied all known systems in nature.

Similar limits can be found for the other electromagnetic observables. In fact, several of the earlier limits are modified when electrical charge is included. Can you show how the size limit changes when electric charge is taken into account? In fact, a dedicated research field is concerned only with the deduction of the most general limits valid in nature.

Challenge 1442 ny

Paradoxes and challenges around the limits

The limits yield a plethora of interesting paradoxes that can be discussed in lectures and student exercises. All paradoxes can be solved by carefully taking into account the combination of effects from general relativity and quantum theory. All apparent violations only appear when one of the two aspects is somehow forgotten. We study a few examples.

- Several black hole limits are of importance to the universe itself. For example, the observed average mass density of the universe is not far from the corresponding limit. Also the lifetime limit is obviously valid for the universe and provides an upper limit for its age. However, the age of the universe is far from that limit by a large factor. In fact, since the universe's size and age still increase, the age limit is pushed further into the future with every second that passes. The universe evolves in a way to escape its own decay.
- The content of a system is not only characterized by its mass and charge, but also by its strangeness, isospin, colour charge, charge and parity. Can you deduce the limits for these quantities?

Challenge 1443 r

- In our discussion of black hole limits, we silently assumed that they interact, like any thermal system, in an incoherent way with the environment. What changes in the results of this section when this condition is dropped? Which limits can be overcome?

Challenge 1444 n

Challenge 1445 e

- Can you find a general method to deduce all limits?
- Brushing some important details aside, we can take the following summary of our study of nature. Galilean physics is that description for which the difference between composed and elementary systems does not exist. Quantum theory is the description of nature with no (really large) composed systems; general relativity is the description of nature with no elementary systems. This distinction leads to the following interesting conclusion. A unified theory of nature has to unify quantum theory and general relativiv-

ity. Since the first theory affirms that no (really large) composed systems exist, while the other that no elementary systems exist, a unified theory should state that *no* systems exist. This strange result indeed seems to be one way to look at the issue. The conclusion is corroborated by the result that in the unified description of nature, the observables time, space and mass cannot be distinguished clearly from each other, which implies that systems cannot be clearly distinguished from their surroundings. To be precise, systems thus do not really exist at unification energy.

Page 985

Limits to measurement precision and their challenge to thought

We now know that in nature, every physical measurement has a lower and an upper bound. One of the bounds is size-dependent, the other is absolute. So far, no violation of these claims is known. The smallest relative measurement error possible in nature is thus the ratio of the two bounds. In short, a smallest length, a highest force and a smallest action, when taken together, imply that *all measurements are limited in precision.*

A fundamental limit to measurement precision is not a new result any more. But it raises many issues. If the mass and the size of any system are themselves imprecise, can you imagine or deduce what happens then to the limit formulae given above?

Challenge 1446 r

Due to the fundamental limits to measurement precision, *the measured values of physical observables do not require the full set of real numbers.* In fact, limited precision implies that no observable can be described by real numbers. We thus recover again a result that appears whenever quantum theory and gravity are brought together.

Ref. 1056

In addition, we found that measurement errors increase when the characteristic measurement energy approaches the Planck energy. In that domain, the measurement errors of any observable are comparable with the measurement values themselves.

Limited measurement precision thus implies that at the Planck energy it is impossible to speak about points, instants, events or dimensionality. Limited measurement precision also implies that at the Planck length it is impossible to distinguish positive and negative time values: particle and anti-particles are thus not clearly distinguished at Planck scales. A smallest length in nature thus implies that there is no way to define the exact boundaries of objects or elementary particles. However, a boundary is what separates matter from vacuum. In short, a minimum measurement error means that, at Planck scales, it is impossible to distinguish objects from vacuum with complete precision. To put it bluntly, at Planck scales, time and space do not exist.

Ref. 1056

The mentioned conclusions are the same as those that are drawn by modern research on unified theories. The force limit, together with the other limits, makes it possible to reach the same conceptual results found by string theory and the various quantum gravity approaches. To show the power of the maximum force limit, we now explore a few conclusions which go beyond present approaches.

Measurement precision and the existence of sets

The impossibility of completely eliminating measurement errors has an additional and important consequence. In physics, it is assumed that nature is a *set* of components. These components are assumed to be separable from each other. This tacit assumption is introduced in three main situations: it is assumed that matter consists of separable particles, that space-time consists of separable events or points, and that the set of states consists of separable initial conditions. So far, all of physics has thus built its complete description of nature on the concept of *set*.

A fundamentally limited measurement precision implies that nature is *not* a set of such separable elements. A limited measurement precision implies that distinguishing physical entities is possible only approximately. The approximate distinction is only possible at energies much lower than the Planck energy. As humans we do live at such smaller energies; thus we can safely make the approximation. Indeed, the approximation is excellent; we do not notice any error when performing it. But the discussion of the situation at Planck energies shows that a perfect separation is impossible in principle. In particular, at the cosmic horizon, at the big bang, and at Planck scales any precise distinction between two events or two particles becomes impossible. Ref. 1056, Ref. 1059

Another way to reach this result is the following. Separation of two entities requires *different measurement results*, such as different positions, different masses, different velocities, etc. Whatever observable is chosen, at the Planck energy the distinction becomes impossible, due to the large measurements errors. Only at everyday energies is a distinction approximately possible. Any distinction between two physical systems, such as between a toothpick and a mountain, is thus possible only *approximately*; at Planck scales, a boundary cannot be drawn. Ref. 1056

A third argument is the following. In order to *count* any entities in nature – a set of particles, a discrete set of points, or any other discrete set of physical observables – the entities have to be separable. The inevitable measurement errors, however, contradict separability. At the Planck energy, it is thus impossible to count physical objects with precision. *Nature has no parts.* Ref. 1056

In short, at Planck energies a perfect separation is impossible in principle. We cannot distinguish observations at Planck energies. In short, *at Planck scale it is impossible to split nature into separate entities.* There are no mathematical elements of any kind – or of any set – in nature. Elements of sets cannot be defined. As a result, in nature, neither discrete nor continuous sets can be constructed. *Nature does not contain sets or elements.* Ref. 1059

Since sets and elements are only approximations, the concept of 'set', which assumes separable elements, is already too specialized to describe nature. Nature cannot be described at Planck scales – i.e., with full precision – if any of the concepts used for its description presupposes sets. However, all concepts used in the past twenty-five centuries to describe nature – space, time, particles, phase space, Hilbert space, Fock space, particle space, loop space or moduli space – are based on sets. They all must be abandoned at Planck energy. *No approach used so far in theoretical physics, not even string theory or the various quantum gravity approaches, satisfies the requirement to abandon sets.* Nature is one and has no parts. *Nature must be described by a mathematical concept that does not contain any set.* This requirement must guide us in the future search for the unification of relativity and quantum theory.

> Es ist fast unmöglich, die Fackel der Wahrheit durch ein Gedränge zu tragen, ohne jemandem den Bart zu sengen.*
>
> Georg Christoph Lichtenberg (1742–1799)

Why are observers needed?

Certain papers on quantum theory give the impression that observers are indispensable for quantum physics. We have debunked this belief already, showing that the observer in quantum theory is mainly a bath with a definite interaction. Often, humans are observers.

At Planck scales, observers also play a role. At Planck scales, quantum theory is mandatory. In these domains, observers must realize an additional requirement: they must function at low energies. Only at low energy, an observer can introduce sets for the description of nature. Introducing observers is thus the same as introducing sets.

To put it in another way, the limits of human observers is that they cannot avoid using sets. However, human observers share this limitation with video recorders, cameras, computers and pencil and paper. Nothing singles out humans in this aspect.

In simple terms, observers are needed to describe nature at Planck scales only in so far as they are and use sets. We should not get too bloated about our own importance.

A solution to Hilbert's sixth problem

In the year 1900, David Hilbert gave a well-known lecture in which he listed twenty-three of the great challenges facing mathematics in the twentieth century. Most problems provided challenges to many mathematicians for decades afterwards. Of the still unsolved ones, Hilbert's sixth problem challenges mathematicians and physicists to find an *axiomatic* treatment of physics. The challenge has remained in the minds of many physicists since that time. Ref. 1075

Since nature does not contain sets, we can deduce that such an axiomatic description of nature does not exist! The reasoning is simple; we only have to look at the axiomatic systems found in mathematics. Axiomatic systems define mathematical structures. These structures are of three main types: algebraic systems, order systems or topological systems. Most mathematical structures – such as symmetry groups, vector spaces, manifolds or fields – are combinations of all three. But all mathematical structures contain sets. Mathematics does not provide axiomatic systems that do not contain sets. The underlying reason is that every mathematical concept contains at least one set. Ref. 1059

Furthermore, all physical concepts used so far in physics contain sets. For humans, it is difficult even simply to *think* without first defining a set of possibilities. However, nature is different; nature does not contain sets. Therefore, *an axiomatic formulation of physics is impossible.* Of course, this conclusion does not rule out unification in itself; however, it does rule out an axiomatic version of it. The result surprises, as separate axiomatic treatments of quantum theory or general relativity (see above) *are* possible. Indeed, only their

* It is almost impossible to carry the torch of truth through a crowd without scorching somebody's beard.

unification, not the separate theories, must be approached without an axiomatic systems. *Axiomatic systems in physics are always approximate.* The requirement to abandon axiomatic systems is one of the reasons for the difficulties in reaching the unified description of nature.

Outlook

Physics can be summarized in a few limit statements. They imply that in nature every physical observable is limited by a value near the Planck value. The speed limit is equivalent to special relativity, the force limit to general relativity, and the action limit to quantum theory. Even though this summary could have been made (or at least conjectured) by Planck, Einstein or the fathers of quantum theory, it is much more recent. The numerical factors for most limit values are new. The limits provoke interesting Gedanken experiments, none of which leads to violations of the limits. On the other hand, the force limit is not yet within direct experimental reach.

The existence of limit values to all observables implies that the description of space-time with a continuous manifold is not correct at Planck scales; it is only an approximation. For the same reason, is predicted that elementary particles are not point-like. Nature's limits also imply the non-distinguishability of matter and vacuum. As a result, the structure of particles and of space-time remains to be clarified. So far, we can conclude that nature can be described by sets only approximately. The limit statements show that Hilbert's sixth problem cannot be solved and that unification requires fresh approaches, taking unbeaten paths into unexplored territory.

We saw that at Planck scales there is no time, no space, and there are no particles. Motion is a low energy phenomenon. Motion only appears for low-energy observers. These
Page 28
are observers who use sets. The (inaccurate) citation of Zeno at the beginning of our walk, stating that motion is an illusion, turns out to be correct! Therefore, we now need to find out how motion actually arises.

The discussion so far hints that motion appears as soon as sets are introduced. To check this hypothesis, we need a description of nature without sets. The only way to avoid the use of sets seems a description of empty space-time, radiation and matter as being made of the same underlying entity. The inclusion of space-time dualities and of interaction dualities is most probably a necessary step. Indeed, both string theory and modern quantum gravity attempt this, though possibly not yet with the full radicalism necessary. Realizing this unification is our next task.

Bibliography

1047 Bohr explained the indivisibilty of the quantum of action in his famous Como lecture. N. Bohr, *Atomtheorie und Naturbeschreibung*, Springer, Berlin, 1931. More statements about the indivisibility of the quantum of action can be found in N. Bohr, *Atomic Physics and Human Knowledge*, Science Editions, New York, 1961. For summaries of Bohr's ideas by others see Max Jammer, *The Philosophy of Quantum Mechanics*, Wiley, first edition, 1974, pp. 90–91, and John Honner, *The Description of Nature – Niels Bohr and the Philosophy of Quantum Physics*, Clarendon Press, 1987, p. 104. Cited on page 986.

1048 For an overview of the quantum of action as basis of quantum theory, see the introduction to quantum theory in this textbook, available as by C. Schiller, An appetizer – quantum theory for poets and lawyers, http://www.motionmountain.net/C-5-QEDA.pdf. Cited on page 987.

1049 The first published statements of the principle of maximum force were in the present textbook and independently in the paper by G.W. Gibbons, The maximum tension principle in general relativity, *Foundations of Physics* **32**, pp. 1891–1901, 2002, or http://www.arxiv.org/abs/hep-th/0210109. Cited on pages 988 and 1000.

1050 Maximal luminosity of often mentioned in studies around gravitational wave detection; nevertheless, the general maximum is never mentioned. See for example L. Ju, D.G. Blair & C. Zhao, Detection of gravitational waves, *Reports on Progress in Physics* **63**, pp. 1317–1427, 2000. See also C. Misner, K. Thorne & J.A. Wheeler, *Gravitation*, Freeman, 1973, page 980. Cited on page 989.

1051 See for example Wolfgang Rindler, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001, p. 70 ff, or Ray d'Inverno *Introducing Einstein's Relativity*, Clarendon Press, 1992, p. 36 ff. Cited on page 990.

1052 T. Jacobson, Thermodynamics of spacetime: the Einstein equation of state, *Physical Review Letters* **75**, pp. 1260–1263, 1995, or preprint http://www.arxiv.org/abs/gr-qc/9504004 Cited on page 990.

1053 C. Schiller, Maximum force: a simple principle encompassing general relativity, part of this text and downloadable as http://www.motionmountain.net/C-2-MAXF.pdf. Cited on page 992.

1054 Indeterminacy relations in general relativity are discussed in C.A. Mead, Possible connection between gravitation and fundamental length, *Physical Review B* **135**, pp. 849–862, 1964. See also P.K. Townsend, Small-scale structure of space-time as the origin of the gravitational constant, *Physical Review D* **15**, pp. 2795–2801, 1977, or the paper by M.-T. Jaekel & S. Renaud, Gravitational quantum limit for length measurement, *Physics Letters A* **185**, pp. 143–148, 1994. Cited on page 992.

1055 Minimal length and minimal time intervals are discussed for example by G. Amelino–Camelia, Limits on the measurability of space-time distances in (the semiclassical approximation of) quantum gravity, *Modern Physics Letters A* **9**, pp. 3415–3422, 1994, and by Y.J. Ng & H. Van Dam, Limit to space-time measurement, *Modern Physics Letters A* **9**, pp. 335–340, 1994. Many other authors have explored the topic. Cited on page 994.

1056 A pedagogical summary is C. Schiller, http://www.motionmountain.net/C10-QMGR.pdf or C. Schiller, Le vide diffère-t-il de la matière? in E. Gunzig & S. Diner, editeurs, *Le vide – Univers du tout et du rien – Des physiciens et des philosophes s'interrogent*, Les Editions de l'Université de Bruxelles, 1998. An older English version is also available as C. Schiller, Does matter differ from vacuum? http://www.arxiv.org/abs/gr-qc/9610066. Cited on pages 994, 995, 998, 1009, and 1010.

1057 Maximal curvature, as well as area and volume quantization are discussed by A. Ashtekar, Quantum geometry and gravity: recent advances, http://www.arxiv.org/abs/gr-qc/0112038 and by A. Ashtekar, Quantum geometry in action: big bang and black holes, preprint downloadable at http://www.arxiv.org/abs/math-ph/0202008. Cited on page 994.

1058 This was kindly pointed out by Achim Kempf. The story is told in A.D. Sakharov, *General Relativity and Gravitation* **32**, p. 365, 2000, a reprint of a paper of 1969. Cited on page 995.

1059 About set definition and Hilbert's sixth problem see C. Schiller, Nature at large scales

– is the universe something or nothing? http://www.motionmountain.net/C11-LGSM.pdf. Cited on pages 995, 998, 1010, and 1011.

1060 Several incorrect counterclaims to the entropy limit were made in R. Bousso, The holographic principle, *Review of Modern Physics* 74, pp. 825–874, 2002, also available as downloadable preprint at http://www.arxiv.org/abs/hep-th/0203101. Bousso has changed position in the meantime. Cited on pages 995 and 1006.

1061 G. Amelino–Camelia, Doubly-special relativity: first results and key open problems, *International Journal of Modern Physics* **11**, pp. 1643–1669, 2002, electronic preprint available as http://www.arxiv.org/abs/gr-qc/0210063. See also J. Magueijo & L. Smolin, Lorentz invariance with an invariant energy scale, *Physical Review Letters* **88**, p. 190403, 2002, or http://www.arxiv.org/abs/hep-th/0112090. Cited on page 995.

1062 Maximons, elementary particles of Planck mass, are discussed by A.D. Sakharov, Vacuum quantum fluctuations in curved space and the theory of gravitation, Soviet Physics – Doklady, **12**, pp. 1040–1041, 1968. Cited on page 996.

1063 Minimal entropy is discussed by L. Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, *Zeitschrift für Physik* **53**, pp. 840–856, 1929. This classic paper can also be found in English translation in his collected works. Cited on page 996.

1064 See for example A.E. Shalyt-Margolin & A.Ya. Tregubovich, Generalized uncertainty relation in thermodynamics, http://www.arxiv.org/abs/gr-qc/0307018, or J. Uffink & J. van Lith-van Dis, Thermodynamic uncertainty relations, *Foundations of Physics* **29**, pp. 655–692, 1999. Cited on page 996.

1065 Gamma ray bursts are discussed by G. Preparata, R. Ruffini & S.-S. Xue, The dyadosphere of black holes and gamma-ray bursts, *Astronomy and Astrophysics* **338**, pp. L87–L90, 1998, and C.L. Bianco, R. Ruffini & S.-S. Xue, The elementary spike produced by a pure e^+e^- pair-electromagnetic pulse from a black hole: the PEM pulse, *Astronomy and Astrophysics* **368**, pp. 377–390, 2001. Cited on page 998.

1066 See for example the review in C.W.J. Beenakker & al., Quantum transport in semiconductor nanostructures, pp. 1–228, in H. Ehrenreich & D. Turnbull, eds., *Solid State Physics*, **44**, Academic Press, Cambridge, 1991. The prediction of a future Nobel prize was made to the authors by C. Schiller in 1992. Cited on page 998.

1067 Brian Greene, *The Elegant Universe - Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*, Vintage, London, 2000, Cited on page 998.

1068 A discussion of a different electric uncertainty relation, namely between current and charge, can be found in Y-Q. Li & B. Chen, Quantum theory for mesoscopic electronic circuits and its applications, http://www.arxiv.org/abs/cond-mat/9907171. Cited on page 999.

1069 Wolfgang Rindler, *Relativity – Special, General and Cosmological*, Oxford University Press, 2001, p. 230. Cited on page 999.

1070 Hans C. Ohanian & Remo Ruffini, *Gravitation and Spacetime*, W.W. Norton & Co., 1994. Cited on page 1002.

1071 A maximal acceleration for microscopic systems was stated by E.R. Caianiello, *Lettere al Nuovo Cimento* **41**, p. 370, 1984. It is discussed by G. Papini, Shadows of a maximal acceleration, http://www.arxiv.org/abs/gr-qc/0211011. Cited on page 1003.

1072 The entropy limit for black holes is discussed by J.D. Bekenstein, Entropy bounds and black hole remnants, *Physical Review D* **49**, pp. 1912–1921, 1994. See also J.D. Bek-

ENSTEIN, Universal upper bound on the entropy-to-energy ration for bounded systems, *Physical Review D* 23, pp. 287–298, 1981. Cited on page 1005.

1073 Private communications with Jos Uffink and Bernard Lavenda. Cited on page 1005.

1074 P. KOVTUN, D.T. SON & A.O. STARINETS, A viscosity bound conjecture, preprint found at http://www.arxiv.org/abs/hep-th/0405231. Cited on page 1006.

1075 H. WUSSING & P.S. ALEXANDROV, (eds.) *Die Hilbertschen Probleme*, Akademische Verlagsgesellschaft Geest & Portig, 1983, or BEN H. YANDELL, *The Honours Class: Hilbert's Problems and their Solvers*, A.K. Peters, 2002. Cited on page 1011.

38. The shape of points – extension in nature

> Nil tam difficile est, quin quaerendo investigari possiet.
>
> Terence*

The expressions for the Compton wavelength $\lambda = h/mc$ and for the Schwarzschild radius $r_s = 2Gm/c^2$ imply a number of arguments which lead to the conclusion that at Planck energies, space-time points and point particles must be described, in contrast to their name, by *extended* entities. These arguments point towards a connection between microscopic and macroscopic scales, confirming the present results of string theory and quantum gravity. At the same time, they provide a pedagogical summary of this aspect of present day theoretical physics.**

1 ▪ It is shown that any experiment trying to measure the size or the shape of an elementary particle with high precision inevitably leads to the result that at least one dimension of the particle is of macroscopic size.

2 ▪ There is no data showing that empty space-time is continuous, but enough data showing that it is not. It is then argued that in order to build up an entity such as the vacuum, that is extended in three dimensions, one necessarily needs extended building blocks

3 ▪ The existence of minimum measurable distances and time intervals is shown to imply the existence of space-time duality, i.e. a symmetry between very large and very small distances. Space-time duality in turn implies that the fundamental entities that make up vacuum and matter are extended.

4 ▪ It is argued that the constituents of the universe and thus of space-time, matter and radiation cannot form a set. As a consequence any precise description of nature must use extended entities.

5 ▪ The Bekenstein–Hawking expression for the entropy of black holes, in particular its surface dependence, confirms that both space-time and particles are composed of extended entities.

6 ▪ Trying to extend statistical properties to Planck scales shows that both particles and space-time points behave as braids at high energies, a fact which also requires extended entities.

7 ▪ The Dirac construction for spin provides a model for fermions, without contradiction with experiments, that points to extended entities.

An overview of other arguments in favour of extended entities provided by present research efforts is also given. To complete the discussion, experimental and theoretical checks for the idea of extended building blocks of nature are presented.

* 'Nothing is so difficult that it could not be investigated.' Terence is Publius Terentius Afer (*c.* 190–159 BCE), important roman poet. He writes this in his play *Heauton Timorumenos*, verse 675.

** This section describes a research topic and as such is *not* a compendium of generally accepted results (yet). It was written between December 2001 and May 2002.

Introduction: vacuum and particles

> Our greatest pretenses are built up not to hide the evil and the ugly in us, but our emptyness. The hardest thing to hide is something that is not there.
>
> Eric Hoffer, *The Passionate State of Mind.*

Only the separation of the description of nature into general relativity and quantum theory allows us to use continuous space-time and point particles as basic entities. When the two theories are united, what we use to call 'point' turns out to have quite counter-intuitive properties. We explore a few of them in the following.

Above, we have given a standard argument showing that points do not exist in nature. The Compton wavelength and the Schwarzschild radius together determine a *minimal* length and a *minimal* time interval in nature. They are given (within a factor of order one, usually omitted in the following) by the Planck length and the Planck time, with the values

Ref. 1076, Ref. 1077

$$\begin{aligned} l_{\mathrm{Pl}} = \sqrt{\hbar G/c^3} &= 1.6 \cdot 10^{-35}\,\mathrm{m} \\ t_{\mathrm{Pl}} = \sqrt{\hbar G/c^5} &= 5.4 \cdot 10^{-44}\,\mathrm{s}\,. \end{aligned} \qquad (775)$$

The existence of a minimal length and space interval in nature implies that points in space, time or space-time have no experimental backing and that we are forced to part from the traditional idea of continuity. Even though, properly speaking, points do not exist, and thus space points, events or point particles do not exist either, we can still ask what happens when we study these entities in detail. The results provide many fascinating surprises.

Page 935

Using a simple *Gedanken experiment*, we have found above that particles and space-time cannot be distinguished from each other at Planck scales. The argument was the following. The largest mass that can be put in a box of size R is a black hole with a Schwarzschild radius of the same value. That is also the largest possible mass measurement error. But any piece of vacuum also has a *smallest* mass measurement error.

The issue of a smallest mass measurement error is so important that it merits special attention. Mass measurement errors always prevent humans to state that a region of space has zero mass. In exactly the same way, also nature cannot 'know' that the mass of a region is zero, provided that this error is due to quantum indeterminacy. Otherwise nature would circumvent quantum theory itself. Energy and mass are always unsharp in quantum theory. We are not used to apply this fact to vacuum itself, but at Planck scales we have to. We remember from quantum field theory that the vacuum, like any other system, has a mass; of course, its value is zero for long time averages. For finite measuring times, the mass value will be uncertain and thus different from zero. Not only limitations in time, but also limitations in space lead to mass indeterminacy for the vacuum. These indeterminacies in turn lead to a minimum mass errors for vacuum regions of finite size. Quantum theory implies that nobody, not even nature, knows the *exact* mass value of a system or of a region of empty space.

A box is empty if it does not contain anything. But emptiness is not well defined for photons with wavelength of the size R of the box or larger. Thus the mass measurement error for an 'empty' box – corresponding to what we call vacuum – is due to the indeterminacy relation and is given by that mass whose Compton wavelength matches the size of

Page 935 Ref. 1076

the box. As shown earlier on, the same mass value is found by every other Gedanken experiment: trying to determine the gravitational mass by weighing the 'piece' of vacuum or by measuring its kinetic energy gives the same result. Another, but in the end equivalent way to show that a region of vacuum has a finite mass error is to study how the vacuum energy depends on the position indeterminacy of the border of the region. Any region is defined through its border. The position indeterminacy of the border will induce a mass error for the contents of the box, in the same way that a time limit does. Again, the resulting mass error value for a region of vacuum is the one for which the box size is the Compton wavelength.

Summarizing, for a box of size R, nature allows only mass values and mass measurement error value m between two limits:

$$\text{(full box)} \quad \frac{c^2 R}{G} \geqslant m \geqslant \frac{\hbar}{cR} \quad \text{(empty box)} . \tag{776}$$

We see directly that for sizes R of the order of the Planck length, the two limits coincide; they both give the Planck mass

$$M_{\mathrm{Pl}} = \frac{\hbar}{c\, l_{\mathrm{Pl}}} = \sqrt{\frac{\hbar c}{G}} \approx 10^{-8}\,\mathrm{kg} \approx 10^{19}\,\mathrm{GeV}/c^2 . \tag{777}$$

In other words, for boxes of Planck size, we cannot distinguish a full box from an empty one. This means that there is no difference between vacuum and matter at Planck scales. Of course, a similar statement holds for the distinction between vacuum and radiation. At Planck scales, vacuum and particles cannot be distinguished.

How else can we show that matter and vacuum cannot be distinguished?

The impossibility to distinguish vacuum from particles is a strong statement. A strong statement needs additional proof.

Mass can also be measured by probing its inertial aspect, i.e. by colliding the unknown mass M with known velocity V with a known probe particle of mass m and momentum $p = mv$. We then have

$$M = \frac{\Delta p}{\Delta V} , \tag{778}$$

where the differences are taken between the values before and after the collision. The error δM of such a measurement is simply given by

$$\frac{\delta M}{M} = \frac{\delta \Delta v}{\Delta v} + \frac{\delta m}{m} + \frac{\delta \Delta V}{\Delta V} . \tag{779}$$

At Planck scales we have $\delta \Delta v / \Delta v \approx 1$, because the velocity error is always, like the velocities themselves, of the order of the speed of light. In other words, at Planck scales the mass measurement error is so large that we cannot determine whether a mass is different from zero: vacuum is indistinguishable from matter.

The same conclusion arises if we take light with a wavelength λ as the probe particle. In this case, expression (778) leads to a mass error

$$\frac{\delta M}{M} = \frac{\delta \Delta\lambda}{\Delta\lambda} + \frac{\delta \Delta V}{\Delta V} . \tag{780}$$

In order that photon scattering can probe Planck dimensions, we need a wavelength of the order of the Planck value; but in this case the first term is approximately unity. Again we find that at Planck scales the energy indeterminacy is always of the *same* size as the energy value to be measured. Measurements cannot distinguish between vanishing and non-vanishing mass M at Planck scales. In fact, this result appears for all methods of mass measurement that can be imagined. At Planck scales, matter cannot be distinguished from vacuum.

Challenge 1447 ny

Incidentally, the same arguments are valid if instead of the mass of matter we measure the energy of radiation. In other words, it is also impossible to distinguish radiation from vacuum at high energies. In short, no type of particle differs from vacuum at high energies.

The indistinguishability of particles and vacuum, together with the existence of minimum space-time intervals, suggest that space, time, radiation and matter are macroscopic approximations of an underlying, common and discrete structure. This structure is often called *quantum geometry*. How do the common constituents of the two aspects of nature look like? We will provide several arguments showing that these constituents are *extended* and *fluctuating*.

Ref. 1078

> Also, die Aufgabe ist nicht zu sehen, was noch nie jemand gesehen hat, sondern über dasjenige was jeder schon gesehen hat zu denken was noch nie jemand gedacht hat.*
>
> Erwin Schrödinger

Argument 1: The size and shape of elementary particles

Size is the length of vacuum taken by an object. This definition comes natural in everyday life, quantum theory and relativity. However, approaching Planck energy, vacuum and matter cannot be distinguished: it is impossible to define the boundary between the two, and thus it is impossible to define the size of an object. As a consequence, every object becomes as extended as the vacuum. Care is therefore required.

What happens if Planck energy is *approached*, advancing step by step to higher energy? Every measurement requires comparison with a standard. A standard is made of matter and comparison is performed using radiation (see Figure 372). Thus any measurement requires to distinguish between matter, radiation and space-time. However, the distinction between matter and radiation is possible only up to the (grand) unification energy, which is about an 800th of the Planck energy. Measurements do not allow us to prove that

* 'Our task is not to see what nobody has ever seen, but to think what nobody has ever thought about that which everybody has seen already.'

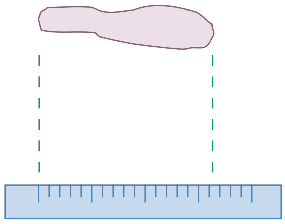


Figure 372 Measurement requires matter and radiation

particles are point-like. Let us take a step back and check whether measurements allow us to say whether particles can at least be contained inside small spheres.

Do boxes exist?

The first and simplest way to determine the size of a compact particle such as a sphere or something akin to it, is to measure the size of a *box* it fits in. To be sure that the particle fits inside, we first of all must be sure that the box is tight. This is done by checking whether something, such as matter or radiation, can leave the box. However, nature does not provide a way to ensure that a box has no holes. Potential hills cannot get higher than the maximum energy, namely the Planck energy. The tunnelling effect cannot be ruled out. In short, there is no way to make fully tight boxes.

In addition, already at the unification energy there is no way to distinguish between the box and the object enclosed in it, as all particles can be transformed from any one type into any other.

Let us cross-check this result. In everyday life, we call particles 'small' because they can be enclosed. Enclosure is possible because in daily life walls are impenetrable. However, walls are impenetrable for matter particles only up to roughly 10 MeV and for photons only up to 10 keV. In fact, boxes do not even exist at medium energies. We thus cannot extend the idea of 'box' to high energies at all.

In summary, we cannot state that particles are compact or of finite size using boxes. We need to try other methods.

Can the Greeks help? – The limits of knifes

The Greeks deduced the existence of atoms by noting that division of matter must end. In contrast, whenever we think of space (or space-time) as made of points, we assume that it can be subdivided without end. Zeno noted this already long time ago and strongly criticized this assumption. He was right: at Planck energy, infinite subdivision is impossible. Any attempt to divide space stops at Planck dimensions at the latest. The process of cutting is the insertion of a wall. Knifes are limited in the same ways that walls are. The limits of walls imply limits to size determination.

In particular, the limits to walls and knives imply that at Planck energies, a cut does not necessarily lead to two separate parts. One cannot state that the two parts have been really separated; a thin connection between can never be excluded. In short, cutting objects at Planck scales does not prove compactness.

Are cross-sections finite?

Particles are not point like. Particles are not compact. Are particles are at least of finite size?

To determine particle size, we can take try to determine their departure from point-likeness. At high energy, detecting this departure requires scattering. For example, we can suspend the particle in some trap and then shoot some probe at it. What happens in a scattering experiment at high energies? The question has been studied already by Leonard Susskind and his coworkers. When shooting at the particle with a high energy probe, the scattering process is characterized by an interaction time. Extremely short interaction times imply sensitivity to the size and shape fluctuations due to the quantum of action. An extremely short interaction time provides a cut-off for high energy shape and size fluctuations and thus determines the measured size. As a result, the size measured for any microscopic, but extended object *increases* when the probe energy is increased towards the Planck value.

Ref. 1079

In summary, even though at experimentally achievable energies the size is always smaller than measurable, when approaching the Planck energy, size increases above all bounds. As a result, at high energies we cannot give a limit to sizes of particles! In other words, since particles are not point-like at everyday energy, at Planck energies they are enormous: particles are extended.

That is quite a deduction. Right at the start of our mountain ascent, we distinguished objects from their environment. Objects are by definition localized, bounded and compact. All objects have a boundary, i.e. a surface which itself does not have a boundary. Objects are also bounded in abstract ways; boundedness is also a property of the symmetries of any object, such as its gauge group. In contrast, the environment is not localized, but extended and unbounded. All these basic assumptions disappear at Planck scales. At Planck energy, it is impossible to determine whether something is bounded or compact. Compactness and localisation are only approximate properties; they are not correct at high energies. The idea of *a point particle is a low energy, approximated concept.*

Particles at Planck scales are as extended as the vacuum. Let us perform another check of this conclusion.

Can one take a photograph of a point?

check the greek

Καιρὸν γνῶθι.*

Pittacus.

Humans or any other types of observers can only observe a part of the world with finite resolution in time and in space. In this, humans resemble a film camera. The highest possible resolution has (almost) been discovered in 1899: the Planck time and the Planck length. No human, no film camera and no measurement apparatus can measure space or time intervals smaller than the Planck values. But what would happen if we took photographs with shutter times approaching the Planck time?

Ref. 1080
Ref. 1076, Ref. 1081

* 'Recognize the right moment.' also rendered as: 'Recognize thine opportunity.' Pittacus (Pittakos) of Mitylene, (*c.* 650–570 BCE) was the Lesbian tyrant that was also one of the ancient seven sages.

Table 75 Effects of various camera shutter times on photographs

Duration	Blur	Observation possibilities
1 h	high	Ability to see faint quasars at night if motion is compensated
1 s	high	Everyday motion is completely blurred
20 ms	lower	Interruption by eyelids; impossibility to see small changes
10 ms	lower	Effective eye/brain shutter time; impossibility to see tennis ball when hitting it
0.25 ms	lower	Shortest commercial photographic camera shutter time; allows to photograph fast cars
1 µs	very low	Ability to photograph flying bullets; requires strong flashlight
c. 10 ps	lowest	Study of molecular processes; ability to photograph flying light pulses; requires laser light to get sufficient illumination
10 fs	higher	Light photography becomes impossible due to wave effects
100 zs	high	X-ray photography becomes impossible; only γ-ray imaging is left over
shorter times	very high	Photographs get darker as illumination gets dimmer; gravitational effects start playing a role
10^{-41} s	highest	imaging makes no sense

Imagine that you have the world's best shutter and that you are taking photographs at increasingly shorter times. Table 75 gives a rough overview of the possibilities. For shorter and shorter shutter times, photographs get darker and darker. Once the shutter time reaches the oscillation time of light, strange things happen: light has no chance to pass undisturbed; signal and noise become impossible to distinguish; in addition, the moving shutter will produce colour shifts. In contrast to our intuition, the picture would get *blurred* at extremely short shutter times. Photography is not only impossible at long but also at short shutter times.

The difficulty of taking photographs is independent of the used wavelength. The limits move, but do not disappear. A short shutter time τ does not allow photons of energy lower than $\hbar/\tau$ to pass undisturbed. The blur is small when shutter times are those of everyday life, but *increases* when shutter times are shortened towards Planck times. As a result, there is no way to detect or confirm the existence of point objects by taking pictures. Points in space, as well as instants of time, are *imagined* concepts; they do not allow a precise description of nature.

At Planck shutter times, only signals with Planck energy can pass through the shutter. Since at these energies matter cannot be distinguished from radiation or from empty space, all objects, light and vacuum look the same. As a result, it becomes impossible to say how nature looks at shortest times.

But the situation is much worse: a Planck shutter does not exist at all, as it would need to be as small as a Planck length. A camera using it could not be built, as lenses do not work at this energy. Not even a *camera obscura* – without any lens – would work, as diffraction effects would make image production impossible. In other words, the idea that at short shutter times, a photograph of nature shows a frozen version of everyday life, like a stopped movie, is completely wrong! Zeno criticized this image already in ancient Greece, in his discussions about motion, though not so clearly as we can do now. Indeed, at a single instant of time nature is not frozen at all.* At short times, nature is blurred and fuzzy. This is also the case for point particles.

In summary, whatever the intrinsic shape of what we call a 'point' might be, we know that being always blurred, it is first of all a cloud. Whatever method to photograph of a point particle is used, it always shows an extended entity. Let us study its shape in more detail.

What is the shape of an electron?

Given that particles are not point-like, they have a shape. How can we determine it? Everyday object shape determination is performed by *touching* the object from all sides. This works with plants, people or machines. It works with molecules, such as water molecules. We can put them (almost) at rest, e.g. in ice, and then scatter small particles off them. Scattering is just a higher energy version of touching. However, scattering cannot determine shapes smaller than the wavelength of the used probes. To determine a size as small as that of an electron, we need the highest energies available. But we already know what happens when approaching Planck scales: the shape of a particle becomes the shape of all the space surrounding it. Shape cannot be determined in this way.

Another method to determine the shape is to build a tight box filled of wax around the system under investigation. We let the wax cool and and observe the hollow part. However, near Planck energies boxes do not exist. We are unable to determine the shape in this way.

A third way to measure shapes is cutting something into pieces and then study the pieces. But *cutting* is just a low-energy version of a scattering process. It does not work at high energies. Since the term 'atom' means 'uncuttable' or 'indivisible', we have just found out that neither atoms nor indivisible particles can exist. Indeed, there is no way to prove this property. Our everyday intuition leads us completely astray at Planck energies.

A fourth way to measure shapes could appear by distinguishing transverse and longitudinal shape, with respect to the direction of motion. However, for transverse shape we get the same issues as for scattering; transverse shape diverges for high energy. To determine longitudinal shape, we need at least two infinitely high potential walls. Again, we already know that this is impossible.

A further, indirect way of measuring shapes is the measurement of the moment of inertia. A finite moment of inertia means a compact, finite shape. However, when the measurement energy is increased, rotation, linear motion and exchange become mixed up. We do not get meaningful results. Ref. 1076

Still another way to determine shapes is to measure the *entropy* of a collection of particles we want to study. This allows to determine the dimensionality and the number of internal degrees of freedom. At high energies, a collection of electrons would become a black hole. We study the issue separately below, but again we find no new information.

Are these arguments water-tight? We assumed three dimensions at all scales, and assumed that the shape of the particle itself is fixed. Maybe these assumptions are not valid

* In fact, a shutter does not exist even at medium energy, as shutters, like walls, stop existing at around 10 MeV.

at Planck scales. Let us check the alternatives. We have already shown above that due to the fundamental measurement limits, the dimensionality of space-time cannot be de-
Ref. 1076 termined at Planck scales. Even if we could build perfect three-dimensional boxes, holes could remain in other dimensions. It does not take long to see that all the arguments against compactness work even if space-time has additional dimensions.

Is the shape of an electron fixed?

Only an object composed of localized entities, such as a house or a molecule, can have a fixed shape. The smaller a system gets, the more quantum fluctuations play a role. Any entity with a finite size, thus also an elementary particle, cannot have a fixed shape. Every Gedanken experiment leading to finite shape also implies that the shape itself fluctuates. But we can say more.

The distinction between particles and environment resides in the idea that particles have *intrinsic* properties. In fact, all intrinsic properties, such as spin, mass, charge and parity, are localized. But we saw that no intrinsic property is measurable or definable at Planck scales. It is impossible to distinguish particles from the environment, as we know already. In addition, at Planck energy particles have all properties the environment also has. In particular, particles are extended.

In short, we cannot prove by experiments that at Planck energy elementary particles are finite in size in all directions. In fact, *all experiments one can think of are compatible with extended particles*, with 'infinite' size. We can also say that particles have *tails*. More precisely, a particle always reaches the borders of the region of space-time under exploration.

Not only are particles extended; in addition, their shape cannot be determined by the methods just explored. The only possibility left over is also suggested by quantum theory: *The shape of particles is fluctuating.*

We note that for radiation particles we reach the same conclusions. The box argument shows that also radiation particles are extended and fluctuating.

In our enthusiasm we have also settled an important detail about elementary particles. We saw above that any particle which is smaller than its own Compton wavelength must
Page 711 be elementary. If it were composite, there would be a lighter component inside it; this lighter particle would have a larger Compton wavelength than the composite particle. This is impossible, since the size of a composite particle must be larger than the Compton wavelength of its components.*

However, an elementary particle *can* have constituents, provided that they are not compact. The difficulties of compact constituents were already described by Sakharov in the
Ref. 1082 1960s. But if the constituents are extended, they do not fall under the argument, as extended entities have no localized mass. As a result, a flying arrow, Zeno's famous example, cannot be said to be at a given position at a given time, if it is made of extended entities. Shortening the observation time towards the Planck time makes an arrow disappear in

* Examples are the neutron, positronium, or the atoms. Note that the argument does not change when the elementary particle itself is unstable, such as the muon. Note also that the possibility that all components be heavier than the composite, which would avoid this argument, does not seem to lead to satisfying physical properties; e.g. it leads to intrinsically unstable composites.

the same cloud that also makes up space-time.*

In summary, only the idea of points leads to problems at Planck scales. If space-time and matter are imagined to be made, at Planck scales, of *extended and fluctuating* entities, all problems disappear. We note directly that for extended entities the requirement of a non-local description is realized. Similarly, the entities being fluctuating, the requirement of a statistical description of vacuum is realized. Finally, the argument forbidding composition of elementary particles is circumvented, as extended entities have no clearly defined mass. Thus the concept of Compton wavelength cannot be defined or applied. Elementary particles can thus have constituents if they are extended. But if the components are extended, how can compact 'point' particles be formed with them? A few options will be studied shortly.

Ref. 1076

Challenge 1448 e

Argument 2: The shape of points in vacuum

> Thus, since there is an impossibility that [finite] quantities are built from contacts and points, it is necessary that there be indivisible material elements and [finite] quantities.
>
> Aristotle, *Of Generation and Corruption*.

Ref. 1084

We are used to think that empty space is made of spatial points. Let us check whether this is true at high energy. At Planck scales no measurement can give zero length, zero mass, zero area or zero volume. There is no way to state that something in nature is a point without contradicting experimental results. In addition, the idea of a point is an extrapolation of what is found in small empty boxes getting smaller and smaller. However, we just saw that at high energies small boxes cannot be said to be empty. In fact, boxes do not exist at all, as they are never tight and do not have impenetrable walls at high energies.

Also the idea of a point as a continuous subdivision of empty space is untenable. At small distances, space cannot be subdivided, as division requires some sort of dividing wall, which does not exist.

Even the idea of repeatedly putting a point between two others cannot be applied. At high energy, it is impossible to say whether a point is exactly on the line connecting the outer two points; and near Planck energy, there is no way to find a point between them at all. In fact, the term 'in between' makes no sense at Planck scales.

We thus find that space points do not exist, in the same way that point particles do not exist. But there is more; space cannot be made of points for additional reasons. Common sense tells us that points need to be kept *apart* somehow, in order to form space. Indeed, mathematicians have a strong argument stating why physical space cannot be made of mathematical points: the properties of mathematical spaces described by the Banach–Tarski paradox are quite different from that of the physical vacuum. The Banach–Tarski paradox states states that a sphere made of mathematical points can be cut into 5 pieces which can be reassembled into two spheres each of the same volume as the original sphere. Mathematically, volume makes no sense. Physically speaking, we can say that the concept of volume does not exist for continuous space; it is only definable if an intrinsic length

Ref. 1076

* Thus at Planck scales there is no quantum Zeno effect any more.

Ref. 1083

exists. This is the case for matter; it must also be the case for vacuum. But any concept with an intrinsic length, also the vacuum, must be described by one or several extended components.* In summary, we need *extended* entities to build up space-time!

Not only is it impossible to generate a volume with mathematical points; it is also impossible to generate exactly three physical dimensions with mathematical points. Mathematics shows that any compact one-dimensional set has as many points as any compact three-dimensional set. And the same is true for any other pair of dimension values. To build up the physical three-dimensional vacuum we need entities which organize their neighbourhood. This cannot be done with purely mathematical points. The fundamental entities must possess some sort of bond forming ability. Bonds are needed to construct or fill three dimensions instead of any other number. Bonds require extended entities. But also a collection of tangled entities extending to the maximum scale of the region under consideration would work perfectly. Of course the precise shape of the fundamental entities is not known at this point in time. In any case we again find that any constituents of physical three-dimensional space must be extended.

In summary, we need extension to define dimensionality and to define volume. We are not surprised. Above we deduced that the constituents of particles are extended. Since vacuum is not distinguishable from matter, we expect the constituents of vacuum to be extended as well. Stated simply, if elementary particles are not point-like, then space-time points cannot be either.

Measuring the void

To check whether space-time constituents are extended, let us perform a few additional Gedanken experiments. First, let us measure the size of a point of space-time. The clearest definition of size is through the cross section. How can we determine the cross section of a point? We can determine the cross section of a piece of vacuum and determine the number of points inside it. From the two determinations we can deduce the cross section of a single point. At Planck energies however, we get a simple result: the cross section of a volume of empty space is depth independent. At Planck energies, vacuum has a surface, but no depth. In other words, at Planck energy we can only state that a Planck layer covers the surface of a volume. We cannot say anything about its interior. One way to picture the result is to say that space points are long tubes.

Another way to determine the size of a point is to count the points found in a given volume of space-time. One approach is to count the possible positions of a point particle in a volume. However, point particles are extended at Planck energies and indistinguishable from vacuum. At Planck energy, the number of points is given by surface area of the volume divided by the Planck area. Again, the surface dependence suggests that particles are long tubes.

Another approach to count the number of points in a volume is to fill a piece of vacuum with point particles.

Ref. 1085 * Imagining the vacuum as a collection of entities with Planck size in all directions, such as spheres, would avoid the Banach–Tarski paradox, but would not allow to deduce the numbers of dimensions of space and time. It would also contradict all other results of this section. We therefore do not explore it further. Challenge 1449 n

What is the maximum number of particles that fits inside a piece of vacuum?

The maximum mass that fits into a piece of vacuum is a black hole. But also in this case, the maximum mass depends only on the *surface* of the given vacuum piece. The maximum mass increases less rapidly than the volume. In other words, the number of points in a volume is only proportional to the surface area of that volume. There is only one solution: vacuum must be made of extended entities crossing the whole volume, independently of the shape of the volume.

Two thousand years ago, the Greek argued that matter must be made of particles because salt can be dissolved in water and because fish can swim through water. Now that we know more about Planck scales, we have to reconsider the argument. Like fish through water, particles can move through vacuum; but since vacuum has no bounds and since it cannot be distinguished from matter, vacuum cannot be made of particles. However, there is another possibility that allows for motion of particles through vacuum: both vacuum and particles can be made of a web of extended entities. Let us study this option in more detail.

Ref. 1086

Argument 3: The large, the small and their connection

> I could be bounded in a nutshell and count myself a king of infinite space.
>
> William Shakespeare, *Hamlet*.

If two observables cannot be distinguished, there is a symmetry transformation connecting them. For example, when switching observation frame, an electric field may change into a magnetic one. A symmetry transformation means that we can change the viewpoint (i.e. the frame of observation) with the consequence that the same observation is described by one quantity from one viewpoint and by the other quantity from the other viewpoint.

When measuring a length at Planck scales it is impossible to say whether we are measuring the length of a piece of vacuum, the Compton wavelength of a body, or the Schwarzschild diameter of a body. For example, the maximum size for an elementary object is its Compton wavelength. The minimum size for an elementary object is its Schwarzschild radius. The actual size of an elementary object is somewhere in between. If we want to measure the size precisely, we have to go to Planck energy: but then all these quantities are the same. In other words, at Planck scales, there is a *symmetry* transformation between Compton wavelength and Schwarzschild radius. In short, *at Planck scales there is a symmetry between mass and inverse mass.*

As a further consequence, at Planck scales there is a symmetry between size and inverse size. *Matter–vacuum indistinguishability means that there is a symmetry between length and inverse length* at Planck energies. This symmetry is called *space-time duality* or *T-duality* in the literature of superstrings.* Space-time duality is a symmetry between situations at scale $n\, l_{\mathrm{Pl}}$ and at scale $f\, l_{\mathrm{Pl}}/n$, or, in other words, between R and $(f\, l_{\mathrm{Pl}})^2/R$, where the experimental number f has a value somewhere between 1 and 1000.

Ref. 1089

* There is also an *S-duality*, which connects large and small coupling constants, and a *U-duality*, which is the combination of S- and T-duality.

Duality is a genuine non-perturbative effect. It does not exist at low energy, since duality automatically also relates energy E and energy $E_{\mathrm{Pl}}^2/E = \hbar c^3/GE$, i.e. it relates energies below and above Planck scale. Duality is a quantum symmetry. It does not exist in everyday life, as Planck's constant appears in its definition. In addition, it is a general relativistic effect, as it includes the gravitational constant and the speed of light. Let us study duality in more detail.

Small is large?

> [Zeno of Elea maintained:] If the existing are many, it is necessary that they are at the same time small and large, so small to have no size, and so large to be without limits.
>
> Simplicius, *Commentary on the Physics of Aristotle*, 140, 34.

Ref. 1087

To explore the consequences of duality, we can compare it to the 2π rotation symmetry in everyday life. Every object in daily life is symmetrical under a full rotation. For the rotation of an observer, angles make sense only as long as they are smaller than 2π. If a rotating observer would insist on distinguishing angles of $0, 2\pi, 4\pi$ etc., he would get a new copy of the universe at each full turn.

Similarly, in nature, scales R and l_{Pl}^2/R cannot be distinguished. Lengths make no sense when they are smaller than l_{Pl}. If however, we insist on using even smaller values and insist on distinguishing them from large ones, we get a new copy of the universe at those small scales. Such an insistence is part of the standard continuum description of motion, where it is assumed that space and time are described by the real numbers, which are defined for arbitrary small intervals. Whenever the (approximate) continuum description with infinite extension is used, the $R \leftrightarrow l_{\mathrm{Pl}}^2/R$ symmetry pops up.

Duality implies that diffeomorphism invariance is only valid at medium scales, not at extremal ones. At extremal scales, quantum theory has to be taken into account in the proper manner. We do not know yet how to do this.

Space-time duality means that introducing lengths smaller than the Planck length (like when one defines space points, which have size zero) means at the same time introducing things with very large ('infinite') value. Space-time duality means that for every small enough sphere the inside equals outside.

Duality means that if a system has a small dimension, it also has a large one. And vice versa. There are thus no small objects in nature. As a result, *space-time duality is consistent with the idea that the basic entities are extended.*

Unification and total symmetry

So far, we have shown that at Planck energy, time and length cannot be distinguished. Duality has shown that mass and inverse mass cannot be distinguished. As a consequence, length, time and mass cannot be distinguished from each other. Since every observable is a combination of length, mass and time, *space-time duality means that there is a symmetry* *between all observables.* We call it the *total symmetry*.*

Total symmetry implies that there are many types of specific dualities, one for each pair of quantities under investigation. Indeed, in string theory, the number of duality types dis-
Page 514
covered is increasing every year. It includes, among others, the famous electric–magnetic
Ref. 1088
duality we first encountered in the chapter on electrodynamics, coupling constant duality, surface–volume duality, space-time duality and many more. All this confirms that there is an enormous symmetry at Planck scales. Similar statements are also well-known right
Ref. 1089
from the beginning of string theory.

Most importantly, total symmetry implies that gravity can be seen as equivalent to all other forces. Space-time duality shows that unification is possible. Physicist have always dreamt about unification. Duality tells us that this dream can indeed be realized.

It may seem that total symmetry is in complete contrast with what was said in the previous section, where we argued that all symmetries are lost at Planck scales. Which result is correct? Obviously, both of them are.

At Planck scales, all low energy symmetries are indeed lost. In fact, all symmetries that imply a *fixed* energy are lost. Duality and its generalizations however, combine both small and large dimensions, or large and small energies. Most symmetries of usual physics, such as gauge, permutation and space-time symmetries, are valid at each fixed energy separately. But nature is not made this way. The precise description of nature requires to take into consideration large and small energies at the same time. In everyday life, we do not do that. Everyday life is a low and fixed energy approximation of nature. For most of the twentieth century, physicists aimed to reach higher and higher energies. We believed that precision increases with increasing energy. But when we combine quantum theory and gravity we are forced to change this approach; to achieve high precision, we must take both high and low energy into account at the same time.*

The large differences in phenomena at low and high energies are the main reason why unification is so difficult. So far, we were used to divide nature along the energy scale. We thought about high energy physics, atomic physics, chemistry, biology, etc. The differences between these sciences is the energy of the processes involved. But now we are not allowed to think in this way any more. We have to take all energies into account at the same time. That is not easy, but we do not have to despair. Important conceptual progress has been achieved in the last decade of the twentieth century. In particular, we now know that we need only *one single concept* for all things which can be measured. Since there is only one concept, there are many ways to study it. We can start from any (low-energy) concept in physics and explore how it looks and behaves when we approach Planck scales. In the present section, we are looking at the concept of point. Obviously, the conclusions must be the same, independently of the concept we start with, be it electric field, spin, or
Challenge 1450 d
any other. Such studies thus provide a check for the results in this section.

* A symmetry between size and Schwarzschild radius, i.e. a symmetry between length and mass, will lead to general relativity. Additionally, at Planck energy there is a symmetry between size and Compton wavelength. In other words, there is a symmetry between length and 1/mass. It means that there is a symmetry between coordinates and wave functions. Note that this is a symmetry between states and observables. It leads to quantum theory.

* Renormalization energy does connect different energies, but not in the correct way; in particular, it does not include duality.

Unification thus implies to think using duality and using concepts which follow from it. In particular, we need to understand what exactly happens to duality when we restrict Challenge 1451 e ourselves to low energy only, as we do in everyday life. This question is left for the next section.

Argument 4: Does nature have parts?

> Pluralitas non est ponenda sine necessitate.*
>
> William of Occam

Another argument, independent of the ones given above, underlines the correctness of a model of nature made of extended entities. Let us take a little broader view. Any concept for which we can distinguish parts is described by a set. We usually describe nature as a set of objects, positions, instants, etc. The most famous set description of nature is the oldest known, given by Democritus: 'The world is made of indivisible particles and void.' This description was extremely successful in the past; there were no discrepancies with observations yet. However, after 2500 years, the conceptual difficulties of this approach are obvious.

We now know that Democritus was wrong, first of all, because vacuum and matter cannot be distinguished at Planck scales. Thus the word 'and' in his sentence is already mistaken. Secondly, due to the existence of minimal scales, the void cannot be made of 'points,' as we usually assume nowadays. Thirdly, the description fails because particles are not compact objects. Fourth, the total symmetry implies that we cannot distinguish parts in nature; nothing can really be distinguished from anything else with complete precision, and thus the particles or points in space making up the naive model of void cannot exist.

In summary, quantum theory and general relativity together show that in nature, *all differences are only approximate*. Nothing can really be distinguished from anything else with complete precision. In other words, there is no way to define a 'part' of nature, neither for matter, nor for space, nor for time, nor for radiation. *Nature cannot be a set.*

The conclusion that nature is not a set does not come as a surprise. We have already encountered another reason to doubt that nature is a set. Whatever definition we use for the term 'particle', Democritus cannot be correct for a purely logical reason. The description he provided is *not complete*. Every description of nature defining nature as a set of parts necessarily misses certain aspects. Most importantly, it misses the *number* of these parts. In particular, the number of particles and the number of dimensions of space-time Page 633 must be specified if we describe nature as made from particles and vacuum. Above we saw that it is rather dangerous to make fun of the famous statement by Arthur Eddington

Ref. 1090 I believe there are 15,747,724,136,275,002,577,605,653,961,181,555,468,044,717,914-

* 'Multitude should not be introduced without necessity.' This famous principle is commonly called *Occam's razor*. William of Ockham (b. 1285/1295 Ockham, d. 1349/50 München), or Occam in the common Latin spelling, was one of the great thinkers of his time. In his famous statement he expresses that only those concepts which are strictly necessary should be introduced to explain observations. It can be seen as the requirement to abandon beliefs when talking about nature. In addition, at this stage of our mountain ascent it has an even more direct interpretation.

> ,527, 116,709,366,231,425,076,185,631,031,296 protons in the universe and the same number of electrons.

In fact, practically all physicists share this belief; usually they either pretend to favour some other number, or worse, they keep the number unspecified. We have seen during our walk that in modern physics many specialized sets are used to describe nature. We have used vector spaces, linear spaces, topological spaces and Hilbert spaces. But very consistently we refrained, like all physicists, from asking about the origin of their sizes (mathematically speaking of their dimensionality or their cardinality). In fact, it is equally unsatisfying to say that the universe contains some specific number of atoms as it is to say that space-time is made of point-like events arranged in 3+1 dimensions. Both statements are about set sizes in the widest sense. In a complete, i.e. in a unified description of nature the number of smallest particles and the number of space-time points must not be added to the description, but must *result* from the description. Only in this case is unification achieved.

Requiring a complete explanation of nature leads to a simple consequence. Any part of nature is by definition smaller than the whole of nature and different from other parts. As a result, any description of nature by a set *cannot* possibly yield the number of particles nor space-time dimensionality. As long as we insist in using space-time or Hilbert spaces for the description of nature, we *cannot* understand the number of dimensions or the number of particles.

Well, that is not too bad, as we know already that nature is *not* made of parts. We know that parts are only approximate concepts. In short, if nature were made of parts, it could not be a unity, or a 'one.' If however, nature is a unity, a one, it cannot have parts.* Nature cannot be separable exactly. It cannot be made of particles.

To sum up, nature cannot be a set. Sets are lists of distinguishable elements. When general relativity and quantum theory are unified, nature shows no elements: *nature stops being a set at Planck scales*. The result confirms and clarifies a discussion we have started in classical physics. There we had discovered that matter objects were defined using space and time, and that space and time were defined using objects. Including the results of quantum theory, this implies that in modern physics particles are defined with the help of the vacuum and the vacuum with particles. That is not a good idea. We have just seen that since the two concepts are not distinguishable from each other, we cannot define them with each other. Everything is the same; in fact, there is no 'every' and no 'thing.' Since nature is not a set, the circular reasoning is dissolved.

Space-time duality also implies that space is not a set. Duality implies that events cannot be distinguished from each other. They thus do not form elements of some space. Ref. 1092 Phil Gibbs has given the name *event symmetry* to this property of nature. This thought-provoking term, even though still containing the term 'event', underlines that it is impossible to use a set to describe space-time.

---

* As a curiosity, practically the same discussion can already be found, in Plato's *Parmenides*, written in the fourth century BCE. There, Plato musically ponders different arguments on whether nature is or can be a *unity* or a *multiplicity*, i.e. a set. It seems that the text is based on the real visit by Parmenides and Zeno in Athens, where they had arrived from their home city Elea, which lies near Naples. Plato does not reach a conclusion. Modern physics however, does. Ref. 1091

In summary, nature cannot be made of vacuum and particles. That is bizarre. People propagating this idea have been persecuted for 2000 years. This happened to the atomists from Democritus to Galileo. Were their battles it all in vain? Let us continue to clarify our thoughts.

Does the universe contain anything?

Stating that the universe contains something implies that we are able to distinguish the universe from its contents. However, we now know that precise distinctions are impossible. If nature is not made of parts, it is wrong to say that the universe *contains* something.

Let us go further. As nothing can be distinguished, we need a description of nature which allows to state that at Planck energies nothing can be distinguished from anything else. For example, it must be impossible to distinguish particles from each other or from the vacuum. There is only one solution: everything, or at least what we call 'everything' in everyday life, must be made of the same single entity. All particles are made of one same 'piece.' Every point in space, every event, every particle and every instant of time must be made of the same single entity.

An amoeba

> A theory of nothing describing everything is better than a theory of everything describing nothing.

We found that parts are approximate concepts. The parts of nature are not strictly smaller than nature itself. As a result, any 'part' must be extended. Let us try to extract more information about the constituents of nature.

The search for a unified theory is the search for a description in which all concepts appearing are only *approximately* parts of the whole. Thus we need an entity Ω, describing nature, which is not a set but which can be approximated by one. This is unusual. We all are convinced very early in our life that we are a *part* of nature. Our senses provide us with this information. We are not used to think otherwise. But now we have to.

Let us eliminate straight away a few options for Ω. One concept without parts is the empty set. Perhaps we need to construct a description of nature from the empty set? We could be inspired by the usual construction of the natural numbers from the empty set. Page 604 However, the empty set makes only sense as the opposite of some full set. That is not the case here. The empty set is not a candidate for Ω.

Another possibility to define approximate parts is to construct them from multiple copies of Ω. But in this way we would introduce a new set through the back door. In addition, new concepts defined in this way would not be approximate.

We need to be more imaginative. How can we describe a whole which has no parts, but which has parts approximately? Let us recapitulate. The world must be described by a single entity, sharing all properties of the world, but which can be approximated into a set of parts. For example, the approximation should yield a set of space points and a set of particles. We also saw that whenever we look at any 'part' of nature without any approximation, we should not be able to distinguish it from the whole world. In other words, composed entities are not always larger than constituents. On the other hand, composed entities must usually appear to be larger than their constituents. For example, space 'points' or 'point' particles are tiny, even though they are only approximations. Which concept without boundaries can be at their origin? Using usual concepts the world is everywhere at the same time; if nature is to be described by a single constituent, this entity must be extended.

The entity has to be a single one, but it must *seem* to be multiple, i.e. it has to be multiple approximately, as nature shows multiple aspects. The entity must be something folded. It must be possible to count the folds, but only approximately. (An analogy is the question of how many tracks are found on an LP or a CD; depending on the point of view, local or global, one gets different answers.) Counting folds would correspond to a length measurement.

The simplest model would be the use of a single entity which is extended, fluctuating, going to infinity and allowing approximate localization, thus allowing approximate definition of parts and points.* In more vivid imagery, nature could be described by some deformable, folded and tangled up entity: a giant, knotted amoeba. An amoeba slides between the fingers whenever one tries to grab a part of it. A perfect amoeba flows around any knife trying to cut it. The only way to hold it would be to grab it in its entirety. However, for an actor himself made of amoeba strands this is impossible. He can only grab it approximately, by catching part of it and approximately blocking it, e.g. using a small hole so that the escape takes a long time.

To sum up, nature is modelled by an entity which is *a single unity* (to eliminate distinguishability), *extended* (to eliminate localizability) and *fluctuating* (to ensure approximate continuity). A far-reaching, fluctuating fold, like an amoeba. The tangled branches of the amoeba allow a definition of length via counting of the folds. In this way, *discreteness* of space, time and particles could also be realized; the quantization of space-time, matter and radiation thus follows. Any flexible and deformable entity is also a perfect candidate for the realization of diffeomorphism invariance, as required by general relativity.

A simple candidate for the extended fold is the image of a fluctuating, flexible *tube*. Counting tubes implies to determine distances or areas. The minimum possible count of one gives the minimum distance, and thus allows us to deduce quantum theory. In fact, we can use as model any object which has flexibility and a small dimension, such as a tube, a thin sheet, a ball chain or a woven collection of rings. These options give the different but probably equivalent models presently explored in simplicial quantum gravity, in Ashtekar's variables and in superstrings.

Argument 5: The entropy of black holes

We are still collecting arguments to determine particle shape. For a completely different way to explore the shape of particles it is useful to study situations where they appear in large numbers. Collections of high numbers of constituents behave differently if they are point-like or extended. In particular, their entropy is different. Studying large-number entropy thus allows to determine component shape. The best approach is to study situations in which large numbers of particles are crammed in a small volume. This leads

Challenge 1452 ny

* Is this the only method to describe nature? Is it possible to find another description, in particular if space and time are not used as background? The answers are unclear at present.

to study the entropy of black holes. A black hole is a body whose gravity is so strong that even light cannot escape. Black holes tell us a lot about the fundamental entities of nature. It is easily deduced from general relativity that any body whose mass m fits inside the so-called Schwarzschild radius

$$r_S = 2Gm/c^2 \tag{781}$$

is a black hole. A black hole can be formed when a whole star collapses under its own weight. A black hole is thus a macroscopic body with a large number of constituents. For black holes, like for every macroscopic body, an entropy can be defined. The entropy S of
Ref. 1093, Ref. 1094 a macroscopic black hole was determined by Bekenstein and Hawking and is given by

$$S = \frac{k}{l_{Pl}^2}\frac{A}{4} \quad \text{or} \quad S = k\frac{4\pi G m^2}{\hbar c} \tag{782}$$

where k is the Boltzmann constant and $A = 4\pi r_S^2$ is the surface of the black hole horizon.
Ref. 1095 This important result has been derived in many different ways. The various derivations also confirm that space-time and matter are equivalent, by showing that the entropy value can be seen both as an entropy of matter and as one of space-time. In the present context, the two main points of interest are that the entropy is *finite*, and that it is *proportional to the area* of the black hole horizon.

In view of the existence of minimum lengths and times, the finiteness of entropy is not surprising any more. A finite black hole entropy confirms the idea that matter is made of a finite number of discrete entities per volume. The existence of an entropy also shows that these entities behave statistically; they fluctuate. In fact, quantum gravity leads to
Ref. 1096 a finite entropy for any object, not only for black holes; Bekenstein has shown that the entropy of any object is always smaller than the entropy of a (certain type of) black hole of the same mass.

The entropy of a black hole is also proportional to its horizon area. Why is this the
Ref. 1097 case? This question is the topic of a stream of publications up to this day.* A simple way to understand the entropy–surface proportionality is to look for other systems in nature with the property that entropy is proportional to system surface instead of system volume. In general, the entropy of any collection of one-dimensional flexible objects, such as poly-
Ref. 1099 mer chains, shows this property. Indeed, the expression for the entropy of a polymer chain made of N monomers, each of length a, whose ends are kept a distance r apart, is given
Ref. 1100 by

$$S(r) = k\frac{3r^2}{2Na^2} \quad \text{for} \quad Na \gg \sqrt{Na} \gg r \; . \tag{783}$$

The formula is derived in a few lines from the properties of a random walk on a lattice, using only two assumptions: the chains are extended, and they have a characteristic internal length a given by the smallest straight segment. Expression (783) is only valid if the polymers are effectively infinite, i.e. if the length Na of the chain and their effective

Ref. 1098 * The result can be derived from quantum statistics alone. However, this derivation does not yield the proportionality coefficient.

average size, the *elongation* $a\sqrt{N}$, are much larger than the radius r of the region of interest; if the chain length is comparable or smaller than the region of interest, one gets the usual extensive entropy, fulfilling $S \sim r^3$. Thus *only flexible extended entities yield a* $S \sim r^2$ *dependence.*

However, there is a difficulty. From the entropy expression of a black hole we deduce that the elongation $a\sqrt{N}$ is given by $a\sqrt{N} \approx l_{\mathrm{Pl}}$; thus it is much smaller than the radius of a general, macroscopic black hole which can have diameters of several kilometres. On the other hand, the formula for long entities is only valid when the chains are longer than the distance r between the end points.

This difficulty disappears once we remember that space near a black hole is strongly curved. All lengths have to be measured in the same coordinate system. It is well known that for an outside observer, any object of finite size falling into a black hole seems to cover the complete horizon for long times (whereas it falls into the hole in its original size for an observer attached to the object). In short, an extended entity can have a proper length of Planck size but still, when seen by an outside observer, be as long as the horizon of the black hole in question. We thus find that black holes are made of extended entities.

Ref. 1079

Another viewpoint can confirm the result. Entropy is (proportional to) the number of yes/no questions needed to know the exact state of the system. This view of black holes has been introduced by Gerard 't Hooft. But if a system is defined by its surface, like a black hole is, its components must be extended.

Finally, imagining black holes as made of extended entities is also consistent with the so-called no-hair theorem: black holes' properties do not depend on what material falls into them, as all matter and radiation particles are made of the same extended components. The final state only depends on the number of entities and on nothing else. In short, the entropy of a black hole is consistent with the idea that it is made of a big tangle of extended entities, fluctuating in shape.

Argument 6: Exchanging space points or particles at Planck scales

We are still collecting arguments for the extension of fundamental entities in nature. Let us focus on their exchange behaviour. We saw above that points in space have to be eliminated in favour of continuous, fluctuating entities common to space, time and matter. Is such a space 'point' or space entity a boson or a fermion? If we exchange two points of empty space in everyday life, nothing happens. Indeed, quantum field theory is based – among others – on the relation

$$[x, y] = xy - yx = 0 \tag{784}$$

between any two points with coordinates x and y, making them bosons. But at Planck scale, due to the existence of minimal distances and areas, this relation is at least changed to

$$[x, y] = l_{\mathrm{Pl}}^2 + \ldots \,. \tag{785}$$

This means that 'points' are neither bosons nor fermions.* They have more complex exchange properties. In fact, the term on the right hand side will be energy dependent, with
Ref. 1076 an effect increasing towards Planck scales. In particular, we saw that gravity implies that double exchange does not lead back to the original situation at Planck scales. Entities following this or similar relations have been studied in mathematics for many decades:
Ref. 1101 braids. In summary, at Planck scales space-time is not made of points, but of braids or some of their generalizations. Thus quantum theory and general relativity taken together
Ref. 1101 again show that vacuum must be made of extended entities.

Particles behave in a similar way. We know that at low, everyday energies, particles of the same type are *identical*. Experiments sensitive to quantum effects show that there is no way to distinguish them: any system of several identical particles obeys permutation symmetry. On the other hand we know that at Planck energy all low-energy symmetries disappear. We also know that, at Planck energy, permutation cannot be carried out, as it implies exchanging positions of two particles. At Planck energy, nothing can be distinguished from vacuum; thus no two entities can be shown to have identical properties. Indeed, no two particles can be shown to be indistinguishable, as they cannot even be shown to be separate.

What happens when we slowly approach Planck energy? At everyday energies, permutation symmetry is defined by commutation or anticommutation relations of any two particle creation operators

$$a^\dagger b^\dagger \pm b^\dagger a^\dagger = 0 \; . \tag{786}$$

At Planck energies this cannot be correct. At those energies, quantum gravity effects appear and modify the right hand side; they add an energy dependent term that is negligible at experimentally accessible energies, but which becomes important at Planck energy. We
Ref. 1076 know from our experience with Planck scales that exchanging particles twice cannot lead back to the original situation, in contrast to everyday life. It is impossible that a double exchange at Planck energy has no effect, because at planck energy such statements are impossible. The simplest extension of the commutation relation (786) satisfying the re-
Ref. 1101 quirement that the right side does not vanish is again *braid symmetry*. Thus Planck scales suggest that particles are also made of extended entities.

Argument 7: The meaning of spin

In the last argument, we will show that even at everyday energy, the extension of particles makes sense. Any particle is a part of the universe. A part is something which is different from anything else. Being 'different' means that exchange has some effect. *Distinction means possibility of exchange.* In other words, any part of the universe is described by its exchange behaviour. Everyday life tells us that exchange can be seen as composed of rotation. In short, distinguishing parts are described by their rotation behaviour. For this reason, for microscopic particles, exchange behaviour is specified by spin. *Spin distinguishes particles from vacuum.**

* The same reasoning destroys the fermionic or Grassmann coordinates used in supersymmetry.

* With a flat (or other) background, it is possible to define a *local* energy–momentum tensor. Thus particles can be defined. Without background, this is not possible, and only global quantities can be defined. Without

We note that volume does not distinguish vacuum from particles; neither does rest mass or charge: there are particles without rest mass or without charge, such as photons. The only candidate observables to distinguish particles from vacuum are spin and momentum. In fact, linear momentum is only a limiting case of angular momentum. We again find that *rotation behaviour is the basic aspect distinguishing particles from vacuum.*

If spin is the central property distinguishing particles from vacuum, finding a model for spin is of central importance. But we do not have to search for long. An well-known model for spin 1/2 is part of physics folklore. Any belt provides an example, as we discussed in detail Page 719 in chapter VI on permutation symmetry. Any localized structure with any number of long tails attached to it – and reaching the border of the region of space under consideration – has the same properties as a spin 1/2 particle. It is a famous exercise to show that such a model, shown in Figure 373, is indeed invariant under 4π rotation but not under 2π rotations, that two such particles get entangled when exchanged, but get untangled when exchanged twice. The model has all properties of spin 1/2 particles, independently of the precise structure of the central region, which remains unknown at this point. The tail model even has the same problems with highly curved space as real spin 1/2 particles have. We explore the idea in detail shortly.

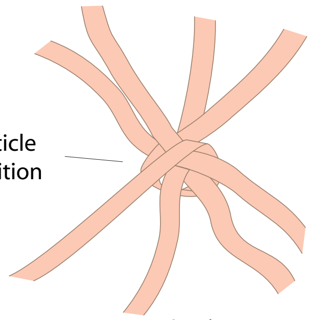


Figure 373 A possible model for a spin 1/2 particle

The tail model thus confirms that rotation is partial exchange. More interestingly, it Ref. 1102 shows that rotation implies connection with the border of space-time. Extended particles can be rotating. Particles can have spin 1/2 provided that they have tails going to the border of space-time. If the tails do not reach the border, the model does not work. Spin 1/2 thus even seems to *require* extension. We again reach the conclusion that extended entities are a good description for particles.

Present research

> To understand is to perceive patterns.
>
> Isaiah Berlin

Ref. 1086 The Greek deduced the existence of atoms because fish can swim through water. They argued that only if water is made of atoms, can a fish find its way through it by pushing the atoms aside. We can ask a similar question when a particle flies through vacuum: why are particles able to move through vacuum at all? Vacuum cannot be a fluid or a solid of small entities, as this would not fix its dimensionality. Only one possibility remains: both vacuum and particles are made of a web of extended entities.

---

background, even particles cannot be defined! Therefore, we assume that we have a slowly varying space-time background in this section.

Describing matter as composed of extended entities is an idea from the 1960s. Describing nature as composed of 'infinitely' extended entities is an idea from the 1980s. Indeed, in addition to the arguments presented so far, present research provides several other approaches that arrive at the same conclusion.

- Bosonization, the construction of fermions using an infinite number of bosons, is a central property of modern unification attempts. It also implies coupling duality, and thus the extension of fundamental constituents. Ref. 1103
- String theory and in particular its generalization to membranes are explicitly based on extended entities, as the name already states. The fundamental entities are indeed assumed to reach the limits of space-time. Ref. 1104
- Research into quantum gravity, in particular the study of spin networks and spin foams, has shown that the vacuum must be thought as a collection of extended entities. Ref. 1105
- In the 1990s, Dirk Kreimer has shown that high-order *QED* diagrams are related to knot theory. He thus proved that extension appears through the back door even when electromagnetism is described using point particles. Ref. 1106
- A recent discovery in particle physics, 'holography', connects the surface and volume of physical systems at high energy. Even if it were not part of string theory, it would still imply extended entities. Ref. 1107
- Other fundamental nonlocalities in the description of nature, such as wave function collapse, can be seen as the result of extended entities. Page 750
- The start of the twenty-first century has brought forwards a number of new approaches, such as string net condensation or knotted particle models. All these make use of extended entities.

Conceptual checks of extension

Is nature really described by extended entities? The idea is taken for granted by all present approaches in theoretical physics.How can we be sure about this result? The arguments presented above provide several possible checks. Challenge 1453 ny

- As Ed Witten likes to say, any unified model of nature must be supersymmetric and dual. The idea of extended entities would be dead as soon as it is shown not to be compatible with these requirements.
- Any model of nature must be easily extendible to a model for black holes. If not, it cannot be correct.
- Showing that the results on quantum gravity, such as the results on the area and volume quantization, are in contradiction with extended entities would directly invalidate the model.
- The same conclusion of extended entities must appear if one starts from *any* physical (low-energy) concept – not only from length measurements – and continues to study how it behaves at Planck scales. If the conclusion were not reached, the idea of extension would not be consistent and thus incorrect.
- Showing that any conclusion of the idea of extension is in contrast with string theory or with M-theory would lead to strong doubts.
- Showing that the measurement of length cannot be related to the counting of folds would invalidate the model.
- Finding a single *Gedanken experiment* invalidating the extended entity idea would

prove it wrong.

Experimental falsification of models based on extended entities

Physics is an experimental science. What kind of data could show that extension is incorrect?

- Observing a single particle in cosmic rays with energy above the Planck energy would bring this approach to a sudden stop. (The present record is a million times lower.)
- Finding any property of nature not consistent with extended entities would spell the end of the idea.
- Finding an elementary particle of spin 0 would invalidate the idea. In particular, finding the Higgs boson and showing that it is elementary, i.e. that its size is smaller than its own Compton wavelength, would invalidate the model.
- Most proposed experimental checks of string theory can also yield information on the ideas presented. For example, Paul Mende has proposed a number of checks on the Ref. 1108 motion of extended objects in space-time. He argues that an extended object moves differently from a mass point; the differences could be noticeable in scattering or dispersion of light near masses.
- In July 2002, the Italian physicist Andrea Gregori has made a surprising prediction valid for any model using extended entities that reach the border of the universe: Ref. 1109 if particles are extended in this way, their mass should depend on the size of the universe. Particle masses should thus change with time, especially around the big bang. This completely new point is still a topic of research.

Incidentally, most of these scenarios would spell the death penalty for almost all present unification attempts.

Possibilities for confirmation of extension models

- The best confirmation would be to find a concrete model for the electron, muon, tau and for their neutrinos. In addition, a concrete model for photons and gravitons is needed. With these models, finding a knot-based definition for the electrical charge and the lepton number would be a big step ahead. All quantum numbers should be topological quantities deduced from these models and should behave as observed.
- Estimating the coupling constants and comparing them with the experimental values is of course the main dream of modern theoretical physics.
- Proving in full detail that extended entities imply exactly three plus one space-time dimensions is still necessary.
- Estimating the total number of particles in the visible universe would provide the final check of any extended entity model.

Generally speaking, the *only* possible confirmations are those from the one-page table Page 887 of unexplained properties of nature given in Chapter X. No other confirmations are possible. The ones mentioned here are the main ones.

Curiosities and fun challenges

> No problem is so formidable that you can't walk away from it.
>
> Charles Schultz

Even though this section already provided sufficient food for thought, here is some more.

▪ If measurements become impossible near Planck energy, we cannot even draw a diagram with an energy axis reaching that value. (See Figure 374) Is this conclusion valid in all cases? Challenge 1454 ny

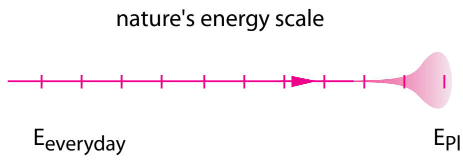


Figure 374 Planck effects make the energy axis an approximation

▪ Quantum theory implies that even if tight walls would exist, the lid of such a box can never be tightly shut. Can you provide the argument? Challenge 1455 n

▪ Is it correct that a detector able to detect Planck mass particles would be of infinite size? What about a detector to detect a particle moving with Planck energy? Challenge 1456 n

▪ Can you provide an argument against the idea of extended entities?* Challenge 1457 e

▪ Does duality imply that the cosmic background fluctuations (at the origin of galaxies and clusters) are the same as vacuum fluctuations? Challenge 1458 ny

▪ Does duality imply that a system with two small masses colliding is the same as one with two large masses gravitating? Challenge 1459 ny

▪ It seems that in all arguments so far we assumed and used continuous time, even though we know it is not. Does this change the conclusions so far? Challenge 1460 d

▪ Duality also implies that large and small masses are equivalent in some sense. A mass m in a radius r is equivalent to a mass m_{Pl}^2/m in a radius l_{Pl}^2/r. In other words, duality transforms mass density from ρ to $\rho_{\mathrm{Pl}}^2/\rho$. Vacuum and maximum density are equivalent. Vacuum is thus dual to black holes.

▪ Duality implies that initial conditions for the big bang make no sense. Duality again shows the uselessness of the idea, as minimal distance did before. As duality implies a symmetry between large and small energies, the big bang itself becomes an unclearly defined concept.

▪ The total symmetry, as well as space-time duality, imply that there is a symmetry between all values an observable can take. Do you agree? Challenge 1461 n

▪ Can supersymmetry be an aspect or special case of total symmetry or is it something else? Challenge 1462 n

▪ Any description is a mapping from nature to mathematics, i.e. from observed differences (and relations) to thought differences (and relations). How can we do this accurately, if differences are only approximate? Is this the end of physics? Challenge 1463 n

▪ What do extended entities imply for the big-bang? Challenge 1464 d

▪ Can you show that going to high energies or selecting a Planck size region of space-time is equivalent to visiting the big-bang? Challenge 1465 d

▪ Additionally, one needs a description for the expansion of the universe in terms of extended entities. First approaches are being explored; no final conclusions can be drawn yet. Can you speculate about the solution? Ref. 1109, Ref. 1110 Challenge 1466 ny

* If so, please email it to the author

An intermediate status report

Wir müssen wissen, wir werden wissen.*
David Hilbert, 1930.

Many efforts for unification advance by digging deeper and deeper into details of quantum field theory and general relativity. Here we took the opposite approach: we took a step back and looked at the general picture. Guided by this idea we found several arguments, all leading to the same conclusion: space-time points and point particles are made of *extended* entities.

Somehow it seems that the universe is best described by a fluctuating, multi-branched entity, a crossing between a giant amoeba and a heap of worms. Another analogy is a big pot of boiling and branched spaghetti. Such an extended model of quantum geometry is beautiful and simple, and these two criteria are often taken as indication, before any experimental tests, of the correctness of a description. We toured topics such as the existence of Planck limits, 3-dimensionality, curvature, renormalization, spin, bosonization, the cosmological constant problem, as well as the search for a background free description of nature. We will study and test specific details of the model in the next section. All these tests concern one of only three possible topics: the construction of space-time, the construction of particles and the construction of the universe. These are the only issues remaining on our mountain ascent of Motion Mountain. We will discuss them in the next section. Before we do so, we enjoy two small thoughts.

Ref. 1111

Sexual preferences in physics

Fluctuating entities can be seen to answer an old and not so serious question. When nature was defined as made of tiny balls moving in vacuum, we described this as a typically male idea. Suggesting that it is male implies that the female part is missing. Which part would that be?

Page 206

From the present point of view, the female part of physics might be the quantum description of the vacuum. The unravelling of the structure of the vacuum, as extended container of localized balls, could be seen as the female half of physics. If women had developed physics, the order of the discoveries would surely have been different. Instead of studying matter, as men did, women might have studied the vacuum first.

Ref. 1112

In any case, the female and the male approaches, taken together, lead us to the description of nature by extended entities. Extended entities, which show that particles are not balls and that the vacuum is not a container, transcend the sexist approaches and lead to the unified description. In a sense, extended entities are thus the politically correct approach to nature.

A physical aphorism

To 'show' that we are not far from the top of Motion Mountain, we give a less serious argument as final curiosity. Salecker and Wigner and then Zimmerman formulated the fundamental limit for the measurement precision τ attainable by a clock of mass M. It is

Ref. 1113, Ref. 1114

* 'We must know, we will know.' This was Hilbert's famous personal credo.

given by $\tau = \sqrt{\hbar T / M c^2}$, where T is the time to be measured. We can then ask what time T can be measured with a precision of a Planck time t_{Pl}, given a clock of the mass of the whole universe. We get a maximum time of

$$T = \frac{t_{\text{Pl}}^2 c^2}{\hbar} M \ . \tag{787}$$

Challenge 1467 e

Inserting numbers, we find rather precisely that the time T is the present age of the universe. With the right dose of humour we can see the result as an omen for the belief that time is now ripe, after so much waiting, to understand the universe down to the Planck scale. We are getting nearer to the top of Motion Mountain. Be prepared for even more fun.

Bibliography

1076 C. Schiller, Does matter differ from vacuum? http://www.arxiv.org/abs/gr-qc/9610066. A French language version has been published as C. Schiller, Le vide diffère-t-il de la matière? in E. Gunzig & S. Diner, editeurs, *Le vide – Univers du tout et du rien – Des physiciens et des philosophes s'interrogent*, Les Éditions de l'Université de Bruxelles, 1998. Cited on pages 1017, 1018, 1021, 1023, 1024, 1025, and 1036.

1077 Most arguments for minimal temporal and spatial intervals are due to M. Schön, Operative time definition and principal uncertainty, preprint http://www.arxiv.org/abs/hep-th/9304024, and to T. Padmanabhan, Limitations on the operational definition of space-time events and quantum gravity, *Classical and Quantum Gravity* 4, pp. L107–L113, 1987; see also his earlier papers referenced there. Cited on page 1017.

1078 A similar point of view, often called monism, was proposed by Baruch Spinoza, *Ethics Demonstrated in Geometrical Order*, 1677, originally in Latin; an affordable French edition is B. Spinoza, *L'Ethique*, Folio-Gallimard, 1954. For a discussion of his ideas, especially his monism, see Don Garret, ed., *The Cambridge Companion to Spinoza*, Cambridge University Press, 1996, or any general text on the history of philosophy. Cited on page 1019.

1079 See L. Susskind & J. Uglum, Black holes, interactions, and strings, http://www.arxiv.org/abs/hep-th/9410074, or L. Susskind, String theory and the principle of black hole complementarity, *Physical Review Letters* 71, pp. 2367–2368, 1993, and M. Karliner, I. Klebanov & L. Susskind, Size and shape of strings, *International Journal of Modern Physics A* 3, pp. 1981–1996, 1988, as well as L. Susskind, Structure of hadrons implied by duality, *Physical Review D* 1, pp. 1182–1186, 1970. Cited on pages 1021 and 1035.

1080 M. Planck, Über irreversible Strahlungsvorgänge, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften zu Berlin, pp. 440–480, 1899. Today it is commonplace to use $\hbar = h/2\pi$ instead of Planck's h, which he originally called b in his paper. Cited on page 1021.

1081 C.A. Mead, Possible connection between gravitation and fundamental length, *Physical Review B* 135, pp. 849–862, 1964. Cited on page 1021.

1082 A.D. Sakharov, Vacuum quantum fluctuations in curved space and the theory of gravitation, *Soviet Physics - Doklady* **12**, pp. 1040–1041, 1968. Cited on page 1024.

1083 P. Facchi & S. Pascazio, Quantum Zeno and inverse quantum Zeno effects, pp. 147–217, in E. Wolf, editor, *Progress in Optics*, **42**, 2001. Cited on page 1025.

1084 Aristotle, *Of Generation and Corruption*, I, II, 316 a 13. See Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 427, 1991. Cited on page 1025.

1085 See for example the speculative model of vacuum as composed of Planck-size spheres proposed by F. Winterberg, *Zeitschrift für Naturforschung* **52a**, p. 183, 1997. Cited on page 1026.

1086 The Greek salt and fish arguments are given by Lucrece, or Titus Lucretius Carus, *De natura rerum*, around 50 BCE. Cited on pages 1027 and 1037.

1087 Cited in Jean-Paul Dumont, *Les écoles présocratiques*, Folio Essais, Gallimard, p. 379, 1991. Cited on page 1028.

1088 D. Olive & C. Montonen, Magnetic monopoles as gauge particles, *Physics Letters* **72B**, pp. 117–120, 1977. Cited on page 1029.

1089 J.H. Schwarz, The second superstring revolution, http://www.arxiv.org/abs/hep-th/9607067. More on dualities in ... Cited on pages 1027 and 1029.

1090 The famous quote is found at the beginning of chapter XI, 'The Physical Universe', in Arthur Eddington, *The Philosophy of Physical Science*, Cambridge, 1939. Cited on page 1030.

1091 Plato, *Parmenides*, *c.* 370 BCE. It has been translated into most languages. Reading it aloud, like a song, is a beautiful experience. A pale reflection of these ideas is Bohm's concept of 'unbroken wholeness'. Cited on page 1031.

1092 P. Gibbs, Event-symmetric physics, http://www.arxiv.org/abs/hep-th/9505089; see also his http://www.weburbia.com/pg/contents.htm website. Cited on page 1031.

1093 J.D. Bekenstein, Black holes and entropy, *Physical Review D* **7**, pp. 2333–2346, 1973. Cited on page 1034.

1094 S.W. Hawking, Particle creation by black holes, *Communications in Mathematical Physics* **43**, pp. 199–220, 1975; see also S.W. Hawking, Black hole thermodynamics, *Physical Review D***13**, pp. 191–197, 1976. Cited on page 1034.

1095 J.B. Hartle & S.W. Hawking, Path integral derivation of black hole radiance, *Physical Review D* **13**, pp. 2188–2203, 1976. See also A. Strominger & C. Vafa, Microscopic origin of Bekenstein–Hawking entropy, *Physics Letters* **B379**, pp. 99–104, 1996, or the downloadable version found at http://www.arxiv.org/abs/hep-th/9601029. For another derivation of black hole entropy, see Gary T. Horowitz & J. Polchinski, A correspondence principle for black holes and strings, *Physical Review D* **55**, pp. 6189–6197, 1997, or http://www.arxiv.org/abs/hep-th/9612146. Cited on page 1034.

1096 J.D. Bekenstein, Entropy bounds and black hole remnants, *Physical Review D* **49**, pp. 1912–1921, 1994. Cited on page 1034.

1097 J. Maddox, When entropy does not seem extensive, *Nature* **365**, p. 103, 1993. Other references on the proportionality of entropy and surface are ... Cited on page 1034.

1098 L. Bombelli, R.K. Koul, J. Lee & R.D. Sorkin, Quantum source of entropy of black holes, *Physical Review D* **34**, pp. 373–383, 1986. Cited on page 1034.

1099 The analogy between polymers and black holes is by G. Weber, Thermodynamics at boundaries, *Nature* **365**, p. 792, 1993. Cited on page 1034.

1100 P.-G. de Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, 1979. Cited on page 1034.

1101 See for example S. Majid, Introduction to braided geometry and q-Minkowski space, preprint found at http://www.arxiv.org/abs/hep-th/9410241 or S. Majid, Duality principle and braided geometry, preprint http://www.arxiv.org/abs/gr-qc/940957. Cited on page 1036.

1102 The relation between spin and statistics has been studied recently by M.V. Berry & J.M. Robbins, Quantum indistinguishability: spin–statistics without relativity or field theory?, in R.C. Hilborn & G.M. Tino, editors, *Spin–Statistics Connection and Commutation Relations*, American Institute of Physics CP545 2000. Cited on page 1037.

1103 An example is given by A.A. Slavnov, Fermi–Bose duality via extra dimension, preprint http://www.arxiv.org/abs/hep-th/9512101. See also the standard work by Michael Stone, editor, *Bosonization*, World Scientific, 1994. Cited on page 1038.

1104 A popular account making the point is the bestseller by the US-american string theorist Brian Greene, *The Elegant Universe – Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*, Jonathan Cape 1999. Cited on page 1038.

1105 The weave model of space-time appears in certain approaches of quantum gravity, as told by A. Ashtekar, Quantum geometry and gravity: recent advances, December 2001, preprint found at http://www.arxiv.org/abs/gr-qc/0112038. See also A. Ashtekar, Quantum geometry in action: big bang and black holes, http://www.arxiv.org/abs/math-ph/0202008. On a slightly different topic, see also S.A. Major, A spin network primer, http://www.arxiv.org/abs/gr-qc/9905020. Cited on page 1038.

1106 A good introduction into his work is the paper D. Kreimer, New mathematical structures in renormalisable quantum field theories, http://www.arxiv.org/abs/hep-th/0211136 to appear in *Annals of Physics* Cited on page 1038.

1107 Introductions to holography are E. Alvarez, J. Conde & L. Hernandez, Rudiments of Holography, http://www.arxiv.org/abs/hep-th/0205075 or R. Bousso, The holographic principle, *Review of Modern Physics* 74, pp. 825–874, 2002, also at http://www.arxiv.org/abs/hep-th/0203101. (However, this otherwise good review has some argumentation errors, as explained on page 812.) The importance of holography in theoretical high energy physics was underlined by the discovery of J. Maldacena, The large N limit of superconformal field theories and supergravity, http://www.arxiv.org/abs/hep-th/9711200. Cited on page 1038.

1108 P.F. Mende, String theory at short distance and the principle of equivalence, preprint available at http://www.arxiv.org/abs/hep-th/9210001. Cited on page 1039.

1109 A. Gregori, Entropy, string theory and our world, http://www.arxiv.org/abs/hep-th/0207195 preprint. Cited on pages 1039 and 1040.

1110 String cosmology is a passtime for many. Examples are N.E. Mavromatos, String cosmology, http://www.arxiv.org/abs/hep-th/0111275 and N.G. Sanchez, New developments in string gravity and string cosmology - a summary report, http://www.arxiv.org/abs/hep-th/0209016. Cited on page 1040.

1111 Searches for background free approaches are described by E. Witten, ... Cited on page 1041.

1112 Not even the most combative feminists have ever said how women-centered physics should look like. This lack of alternatives pervades also the agressive text by Margaret Wertheim, *Pythagoras' Trousers – God, Physics and the Gender Wars*, Fourth Estate, 1997, and thus makes its author look rather silly. Cited on page 1041.

1113 H. Salecker & E.P. Wigner, Quantum limitations of the measurement of space-time distances, *Physical Review* **109**, pp. 571–577, 1958. Cited on page 1041.

1114 E.J. Zimmerman, The macroscopic nature of space-time, *American Journal of Physics* **30**, pp. 97–105, 1962. Cited on page 1041.

1115 See for example C. Rovelli, Notes for a brief history of quantum gravity, electronic preprint available at http://www.arxiv.org/abs/gr-qc/0006061. No citations.

1116 See the http://www.nuclecu.unam.mx/~alberto/physics/stringrev.html website for an up-to date list of string theory and M theory reviews. No citations.

1117 W.G. Unruh, Notes on black hole evaporation, *Physical Review D* **14**, pp. 870–875, 1976. W.G. Unruh & R.M. Wald, What happens when an accelerating observer detects a Rindler particle, *Physical Review D* **29**, pp. 1047–1056, 1984. No citations.

1118 The most prominent proposer of the idea that particles might be knots was William Thomson. See W. Thomson, On vortex motion, *Transactions of the Royal Society in Edinburgh* pp. 217–260, 1868. This paper stimulated work on knot theory, as it proposed the idea that different particles might be differently knotted vortices in the aether. No citations.

Chapter XII

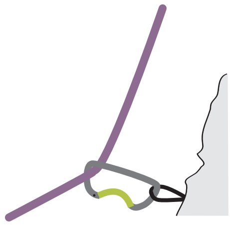
Extension and Unification (Not yet Available)

– CS – this chapter will be made available in the future – CS –

CHAPTER XIII

THE TOP OF THE MOUNTAIN (NOT YET AVAILABLE)

– CS – this chapter will be made available in the future – CS –

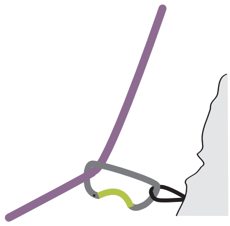


Appendices

Where necessary reference information
for mountain ascents is provided,
preparing the reader for this
and any other future adventure.

Appendix A

Notation and Conventions



Newly introduced and defined concepts in this text are indicated by *italic typeface*. New definitions can also be found in the index, referred to by italic page numbers. Throughout the text SI units are used; they are defined in Appendix B. Experimental results are cited with limited precision, usually only two digits, as this is almost always sufficient. High-precision reference values can be found in Appendices B and C.

In relativity we use the time convention, where the metric has the signature $(+---)$. Ref. 1119 It is used by about 70 % of the literature worldwide. We use indices i, j, k for three-vectors and indices a, b, c, etc. for four-vectors. Other conventions specific to general relativity are explained in the corresponding chapter.

The symbols used in the text

> To avoide the tediouse repetition of these woordes: is equalle to: I will sette as I doe often in woorke use, a paire of paralleles, or Gemowe lines of one lengthe, thus: = , bicause noe .2. thynges, can be moare equalle.
>
> Robert Recorde*

Books are collections of symbols. Most symbols have been developed over hundreds of years; only the clearest and simplest are now in use. In this mountain ascent, the symbols used as abbreviations for *physical* quantities are all taken from the Latin or Greek alphabets and are always defined in the context where they are used. The symbols designating units, constants and particles are defined in Appendices B and C. Just as a note, there is an international standard for them (ISO 31), but the standard is virtually inaccessible; the Ref. 1121 symbols used in this text are those in common use.

The *mathematical* symbols used in this text, in particular those for operations and relations, are given in the following list, together with their origin. The details of their Ref. 1120 history have been extensively studied in the literature.

* Robert Recorde (*c.* 1510–1558), English mathematician and physician; he died in prison, though not for Ref. 1120 his false pretence to be the inventor of the equal sign, which he simply took from his Italian colleagues, but for a smaller crime, namely debt. The quotation is from his *The Whetstone of Witte*, 1557. An image of the quote can be found at the http://members.aol.com/jeff94100/witte.jpg web site.

$+, -$	plus, minus; the plus sign is derived from Latin 'et'	German mathematicians, end of fifteenth century
$\sqrt{}$	read as 'square root'; the sign stems from a deformation of the letter 'r', initial of the Latin 'radix'	used by K. Rudolff in 1525
$=$	equal to	Italian mathematicians, early sixteenth century, then brought to England by R. Recorde
$\{\}, [\,], (\,)$	grouping symbols	use starts in the sixteenth century
$>, <$	larger than, smaller than	T. Harriot 1631
$\times$	multiplied with, times	W. Oughtred 1631
$:$	divided by	G. Leibniz 1684
$\cdot$	multiplied with, times	G. Leibniz 1698
a^n	power	R. Descartes 1637
x, y, z	coordinates, unknowns	R. Descartes 1637
$ax + by + c = 0$	constants and equations for unknowns	R. Descartes 1637
$\mathrm{d}/\mathrm{d}x$, d^2x, $\int y\,\mathrm{d}x$	derivative, differential, integral	G. Leibniz 1675
φx	function of x	J. Bernoulli 1718
$fx, f(x)$	function of x	L. Euler 1734
$\Delta x, \sum$	difference, sum	L. Euler 1755
$\neq$	is different from	L. Euler eighteenth century
$\partial/\partial x$	partial derivative, read like 'd/dx'	it was deduced from a cursive form of 'd' or of the letter 'dey' of the Cyrillic alphabet by A. Legendre in 1786
Δ	Laplace operator	R. Murphy 1833
$\|x\|$	absolute value	K. Weierstrass 1841
∇	read as 'nabla'	introduced by William Hamilton in 1853 and P.G. Tait in 1867, named after the shape of an old Egyptian musical instrument
$[x]$	the measurement unit of a quantity x	twentieth century
∞	infinity	J. Wallis 1655
π	4 arctan 1	H. Jones 1706
e	$\sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n\to\infty}(1 + 1/n)^n$	L. Euler 1736
i	$+\sqrt{-1}$	L. Euler 1777
$\cup, \cap$	set union and intersection	G. Peano 1888
$\in$	element of	G. Peano 1888
$\varnothing$	empty set	André Weil as member of the N. Bourbaki group in the early twentieth century
$\langle\psi\|, \|\psi\rangle$	bra and ket state vectors	Paul Dirac 1930
$\otimes$	dyadic product or tensor product or outer product	unknown

Other signs used here have more complicated origins. The & sign is a contraction of Latin ‘et’ meaning ‘and’, as often is more clearly visible in its variations, such as *&*, the common italic form. Challenge 1468 ny

The punctuation signs used in sentences with modern Latin alphabets, such as , . ; : ! ? ‘ ’ » « – ( ) ... , each have their own history. Many are from ancient Greece, but the question mark is from the court of Charlemagne and exclamation marks appear first in the sixteenth century.* The or *at-sign* probably stems from a medieval abbreviation of Latin ad, meaning ‘at’, in a similar way as the & sign evolved from latin et. In recent years, the *smiley* :-) and its variations have become popular. The smiley is in fact a new edition of the ‘point of irony’ which had been proposed already, without success, by A. de Brahm (1868–1942). Ref. 1122 Ref. 1123

The section sign § dates from the thirteenth century in northern Italy, as was shown by the German palaeographer Paul Lehmann. It was derived from ornamental versions of the capital letter C for ‘capitulum’, i.e. ‘little head’ or ‘chapter.’ The sign appeared first in legal texts, where it is still used today, and then spread also into other domains. Ref. 1124

The paragraph sign ¶ was derived from a simpler ancient form looking like the Greek letter Γ, a sign which was used in manuscripts from ancient Greece until way into the middle ages to mark the start of a new text paragraph. In the middle ages it took the modern form because probably a letter c for ‘caput’ was added in front of it.

Ref. 1125 One of the most important signs of all, the *white space* separating words, was due to Celtic and Germanic influences when these people started using the Latin alphabet. It became commonplace only between the ninth and the thirteenth century, depending on the language in question.

The Latin alphabet

> What is written without effort is in general read without pleasure.
>
> Samuel Johnson (1709–1784)

This text is written using the Latin alphabet. At first sight, this seems to imply that its pronunciation *cannot* be explained in print, in contrast to that of any additional alphabet or to the International Phonetic Alphabet (or IPA). But that is not correct. Physics beats false logic. It is obviously possible to write a text that describes exactly how to move lips, mouth and tongue for each letter, using physical concepts wherever they are necessary. The definitions of pronunciations found in dictionaries make indirect use of this method by referring to the memory of pronounced words or to sounds found in nature.

Historically, the Latin alphabet was derived from the Etruscan, which itself was a derivation of the Greek alphabet. The main forms are

* On the parenthesis see the beautiful book by J. LENNARD, *But I Digress*, Oxford University Press, 1991.

The *ancient* Latin alphabet,
used from the sixth century BCE onwards:

A B C D E F Z H I K L M N O P Q R S T V X

The *classical* Latin alphabet,
used from the second century BCE until the eleventh century:

A B C D E F G H I K L M N O P Q R S T V X Y Z

The letter G was added in the third century BCE by the first Roman to run a fee paying school, Spurius Carvilius Ruga, by adding a horizontal bar to the letter C and substituting the letter Z, which was not used in Latin any more. In the second century BCE, after the conquest of Greece, the Romans included the letters Y and Z from the Greek alphabet at the end of their own (therefore effectively reintroducing the Z) in order to be able to write Greek words. This classical Latin alphabet was stable throughout the next one thousand years.*

The classical Latin alphabet was spread around Europe, Africa and Asia by the Romans during their conquests; due to its simplicity it was adopted by numerous modern languages. Most modern 'Latin' alphabets usually include other letters. The letter W was introduced in the eleventh century in French and was then adopted in most other languages. The letters J and U were introduced in the sixteenth century in Italy, to distinguish them from I and V, which used to have both meanings. In other languages they are used for other sounds. The contractions æ and œ are from the middle ages. Other Latin alphabets include more letters, such as the German *sharp s*, written ß, a contraction of
Ref. 1126 'ss' or 'sz', the nordic letters *thorn*, written Þ or þ, and *eth*, written Ð or ð, taken from the Futhark,** and other signs.

Similarly, lower case letters are not classical Latin; they date only from the middle ages, from the time of Charlemagne. Like most accents such as ê, ç or ä, which were also defined in the middle ages, they were introduced to save the then expensive paper surface by shortening printed words.

> Outside a dog, a book is a man's best friend. Inside a dog, it's too dark to read.
>
> Groucho Marx

Ref. 1127

The Greek alphabet

The Latin alphabet was derived from the Etruscan, the Etruscan from the Greek. The Greek alphabet in turn was derived from the Phoenician or a similar northern Semitic alphabet in the tenth century BCE. In contrast to the Etruscan and Latin alphabets, each

* To meet latin speakers and writers, go to http://www.alcuinus.net/.

** The Runic script or *Futhark*, a type of alphabet used in the middle ages in Germanic, Anglo–Saxon and Nordic countries, probably also derives from the Etruscan alphabet. As the name says, the first letters were f, u, th, a, r, k (in other regions f, u, th, o, r, c). The third letter is the letter thorn mentioned above; it is often written 'Y' in old English, as in 'Ye Olde Shoppe.'

Table 76 The Greek alphabet

Anc.	Clas.		Name	Corr.	
Α	Α	α	alpha	a	1
Β	Β	β	beta	b	2
Γ	Γ	γ	gamma	g, n[1]	3
Δ	Δ	δ	delta	d	4
Ε	Ε	ε	epsilon	e	5
Ϝ ϝ, ϛ			digamma, stigma[2]	w	6
Ζ	Ζ	ζ	zeta	z	7
Η	Η	η	eta	e	8
Θ	Θ	θ	theta	th	9
Ι	Ι	ι	iota	i, j	10
Κ	Κ	κ	kappa	k	20
Λ	Λ	λ	lambda	l	30
Μ	Μ	μ	mu	m	40
Ν	Ν	ν	nu	n	50
Ξ	Ξ	ξ	xi	x	60
Ο	Ο	ο	omicron	o	70
Π	Π	π	pi	p	80
Ϙ ϙ, ϟ			qoppa[3]	q	90
Ρ	Ρ	ρ	rho	r, rh	100
Σ	Σ	σ, ς	sigma[4]	s	200
Τ	Τ	τ	tau	t	300
	Υ	υ	upsilon	y, u[5]	400
	Φ	φ	phi	ph, f	500
	Χ	χ	chi	ch	600
	Ψ	ψ	psi	ps	700
	Ω	ω	omega	o	800
Λ ϡ			sampi[6]	s	900

The regional archaic letters yot, sha and san are not included in the table. The letter san was the ancestor of sampi.
1. Only if before velars, i.e. before kappa, gamma, xi and chi.
2. ‘Digamma’ is the name used for the F-shaped form. It was mainly used as a letter (but also sometimes, in its lower case form, as number), whereas the shape and name ‘stigma’ is used only as number. Both names were deduced from the respective shapes; in fact, the stigma is a medieval, uncial version of the digamma. The name ‘stigma’ is derived from the fact that the letter looks like a sigma with a tau attached under it – though unfortunately not in all modern fonts. The original letter name, also giving its pronunciation, was ‘waw’.
3. The version of qoppa that looks like a reversed and rotated z is still in occasional use in modern Greek. Unicode calls this version ‘koppa’.
4. The second variant of sigma is used only at the end of words.
5. Only if second letter in diphthongs.
6. In older times, the letter sampi was positioned between pi and qoppa.

Greek letter has a proper name, as was the case for the Phoenician alphabet and many of its derivatives. The Greek letter names of course are the origin of the term *alphabet* itself.

In the tenth century BCE, the Ionian or *ancient* (eastern) Greek alphabet consisted of the upper case letters only. In the sixth century BCE several letters were dropped, a few new ones and the lower case versions were added, giving the *classic* Greek alphabet. Still later, accents, subscripts and the breathings were introduced. The following table also gives the values the letters took when they were used as numbers. For this special use the obsolete ancient letters were kept also during the classical period; thus they also have a lower case form.

The Latin correspondence in the table is the standard classical one, used in writing of Greek words. The question about the correct *pronunciation* of Greek has been hotly debated in specialist circles; the traditional *Erasmian* pronunciation does not correspond to the results of linguistic research, nor to the modern Greek one. In classic Greek, the sound that sheep make was βη–βη. (Erasmian pronunciation wrongly insists on a narrow η; modern Greek pronunciation is different for β, which is now pronounced 'v', and for η, which is now pronounced 'i:'.) Obviously, the pronunciation of Greek varied from region to region and over time. For Attic, the main dialect spoken in the classical period, the question is now settled. Linguistic research showed that chi, phi and theta were less aspirated than usually pronounced and sounded like the initials of 'cat', 'perfect' and 'tin'; moreover, the zeta seems to have been pronounced more like 'zd' as in 'buzzed'. For the vowels, contrary to tradition, epsilon is closed and short whereas eta is open and long, omicron is closed and short, whereas omega is wide and long, and upsilon is really a 'u' sound like in 'boot', not a French 'u' or German 'ü.'

The Greek vowels can have rough or smooth *breathings*, *subscripts*, as well as acute, grave, circumflex or diaeresis *accents*. Breathings, used also on ρ, determine whether the letter is aspirated. Accents, interpreted only as stresses in the Erasmian pronunciation, actually represented pitches. Classical Greek could have up to three added signs per letter; modern Greek never has more than one accent.

A descendant of the Greek alphabet* is the *Cyrillic alphabet*, used with slight variations in many slavic languages, such as Russian. However, there exists no standard transcription from Cyrillic to Latin, so that often the same author is spelled differently in different countries and even on different occasions. Ref. 1128

The Hebrew alphabet and other scripts

The phoenician alphabet is also at the origin of the Hebrew consonant alphabet or abjad. Its first letters are given in Table 77.

Table 77 The beginning of the Hebrew abjad

Letter	Name	Corr.	Num.
א	aleph	a	1
ב	beth	b	2
ג	gimel	g	3
ד	daleth	d	4
etc.			

Only the letter aleph is commonly used in mathematics, though others have been proposed. Ref. 1121

* The Greek alphabet is also at the origin of the *Gothic alphabet*, which was defined in the fourth century by Wulfila for the Gothic language, using also a few signs from the Latin and Futhark scripts.

The Gothic alphabet is not to be confused with the so-called *Gothic letters*, a style of the *Latin* alphabet used all over Europe from the eleventh century onwards. In Latin countries, Gothic letters were replaced in the sixteenth century by the *Antiqua*, the ancestor of the type in which this text is set. In other countries, Gothic letters remained in use for much longer. They were used in type and handwriting in Germany until in 1941 the national-socialist government suddenly abolished them. They remain in sporadic use across Europe. In many physics and mathematics books, gothic letters are used to denote vector quantities.

A number of additional writing system are used throughout the world. Experts classify them into five groups. *Phonemic alphabets*, such as Latin or Greek, have a sign for each consonant and vowel. *Abjads* or consonant alphabets, such as Hebrew or Arabic, have a sign for each consonant (sometimes including some vowels), and do not write (most) vowels; most abjads are written from right to left. *Abugidas*, also called syllabic alphabets or alphasyllabaries, such as Balinese, Burmese, Devanagari, Tagalog, Thai, Tibetan or Lao, write consonants and vowels; consonants have an inherent vowel which can be changed into the others by diacritics. *Syllabaries*, such as Hiragana or Ethiopic, have a sign for each syllable of the language. Finally, complex scripts, such as Chinese, Mayan or the Hieroglyphs, use signs which have both sound and meaning. Writing systems can write from right to left, from top to bottom, and can count book pages in the opposite sense to this one.

Page 602

Even though there are about 7000 languages on Earth, there are only about *one hundred* writing systems. About fifty others are not in use any more. * For physical and mathematical formulae though, the sign system presented here, based on Latin and Greek letters, written from left to right and from top to bottom, is a standard the world over. It is used independently of the writing system of the text containing it.

Digits and numbers

Both the digits and the method used in this text to write numbers stem from India. They were brought to the Mediterranean by Arabic mathematicians in the middle ages. The number system used in this text is thus much younger than the alphabet.** The Indian numbers were made popular in Europe by Leonardo of Pisa, called Fibonacci,*** in his book *Liber Abaci* or 'Book of Calculation', which he published in 1202. From that day on mathematics changed. Everybody with paper and pen (the pencil had not been invented yet) was now able to calculate and write down numbers as large as reason allows, or even larger, and to perform calculations with them. The book started:

> Novem figure indorum he sunt 9 8 7 6 5 4 3 2 1. Cum his itaque novem figuris, et cum hoc signo 0, quod arabice zephirum appellatur, scribitur quilibet numerus, ut inferius demonstratur.****

* A well-designed website on the topic is http://www.omniglot.com. The main present and past writing systems are encoded in the unicode standard, which at present contains 52 writing systems. See http://www.unicode.org.

** The story of the development of the numbers is told most interestingly by G. Ifrah, *Histoire universelle des chiffres*, Seghers, 1981, which has been translated into several languages. He sums up the genealogy in ten beautiful tables, one for each digit, at the end of the book. However, the book contains many factual errors, as explained in the http://www.ams.org/notices/200201/rev-dauben.pdf and http://www.ams.org/notices/200202/rev-dauben.pdf review.

It is not correct to call the digits 0 to 9 *Arabic*. Both the actual Arabic digits and the one used in Latin texts such as this one derive from the Indian digits. Only the digits 0, 2, 3 and 7 resemble those used in Arabic writing, provided they are turned clockwise by 90°.

*** Leonardo di Pisa, called Fibonacci (b. *c.* 1175 Pisa, d. 1250 Pisa), Italian mathematician, and the most important mathematician of his time.

**** 'The nine figures of the Indians are: 9 8 7 6 5 4 3 2 1. With these nine figures, and with this sign 0 which in Arabic is called zephirum, any number can be written, as will be demonstrated below.'

The Indian method of writing numbers consists of a large innovation, the *positional system*, and a small one, the digit zero. The positional system, as described by Fibonacci, was so much more efficient to write numbers that it completely replaced the previous *Roman number system*, which writes 1996 as IVMM or MCMIVC or MCMXCVI, as well as the *Greek number system*, in which the Greek letters were used for numbers in the way shown above, thus writing 1996 as ͵αϡϙϛʹ. Compared to these systems, the Indian numbers are a much better technology. Indeed, the Indian system is so practical that calculations done on paper completely eliminated calculations with the help of the *abacus*, which therefore fell in disuse. The abacus is still in use only in those countries which do not use a positional system to write numbers. (The Indian system also eliminated systems to represent numbers with fingers; such systems, which could show numbers up to 10000 and more, have left only one trace: the term 'digit' itself, which derives from the Latin word for finger.) Similarly, only the positional number system allows mental calculations and made and still makes calculating prodigies possible.*

Calendars

The many ways to keep track of time differ greatly from civilization to civilization. The most common calendar, the one used in this text, is at the same time one of the most absurd, as it is a compromise between various political forces who tried to shape it.

In ancient times, independent localized entities, such as tribes or cities, preferred *lunar* calendars, because lunar time keeping is easily organized locally. This led to the use of the month as calendar unit. Centralized states imposed *solar* calendars, based on the year. Solar systems require astronomers and thus a central authority to finance them. For various reasons, farmers, politicians, tax collectors, astronomers and some, but not all, religious groups wanted the calendar to follow the solar year as precisely as possible. The compromises necessary between months and years are the origin of leap days. The compromises required that different months in a year have different lengths; in addition, their lengths are different in different calendars. The most commonly used year–month structure was organized over 2000 years ago by Gaius Julius Ceasar, and is thus called the *Julian calendar*.

The week is an invention of Babylonia, from where it was taken over and spread through the world by various religious groups. (The way astrology and astronomy cooperated to determine the order of the weekdays is explained in the section on gravitation.) Ref. 1129 Even though about three thousand years old, the week was fully included into the Julian calendar only around the year 300, towards the end of the western Roman empire. The final change in the Julian calendar took place between 1582 and 1917 (depending on the country), when more precise measurements of the solar year were used to set a new method to determine leap days, a method still in use today. Together with a reset of the date and the fixation of the week rhythm, this standard is called the *Gregorian calendar* or simply the *modern calendar*. It is used by a majority of the world's population.

* The presently shortest time for finding the thirteenth (integer) root of a hundred digit number, a result with 8 digits, is 11.8 seconds. About the stories and the methods of calculating prodigies, see the fascinating book by Steven B. Smith, *The Great Mental Calculators – The Psychology, Methods and Lives of the Calculating Prodigies*, Columbia University Press, 1983. The books also presents the techniques that they use and that anybody else can use to emulate them.

Despite this complexity, the modern calendar does allow you to determine the day of the week of a given date in your head. Just execute the following 6 steps:

1. take the last two digits of the year, divide by 4, discarding any fraction,
2. add the last two digits of the year,
3. subtract 1 for January or February of a leap year,
4. add 6 for 2000's or 1600's, 4 for 1700's or 2100's,
   2 for 1800's and 2200's, and 0 for 1900's or 1500's,
5. add the day of the month,
6. add the month key value, namely 144 025 036 146 for JFM AMJ JAS OND.

The remainder after division by 7 gives the day of the week, with the correspondence 1-2-3-4-5-6-0 meaning sunday-monday-tuesday-wednesday-thursday-friday-saturday.*

When to start the counting years is a matter of choice. The oldest method not attached to political power structures was the method used in ancient Greece, when years were counted in function of the Olympic games. In those times, people used to say e.g. that they were born in the first year of the twentythird olympiad. Later, political powers always imposed counting years from some important event onwards.** Maybe reintroducing the Olympic counting is worth considering?

Abbreviations and eponyms or concepts?

The scourge of modern physics are sentences like the following:

> The EPR paradox in the Bohm formulation can perhaps be resolved using the GRW approach, using the WKB approximation of the Schrödinger equation.

---

* Remembering the intermediate result for the current year can simplify things even more, especially since the dates 4.4., 6.6., 8.8., 10.10., 12.12., 9.5., 5.9., 7.11., 11.7. and the last day of February all fall on the same day of the week, namely on the year's intermediate result plus 4.

** The present counting of year was defined in the middle ages by setting the date for the foundation of Rome to the year 753 BCE, or 753 *Before Common Era*, and then counting backwards, implying that the BCE years behave almost like negative numbers. However, in this definition the year 1 follows directly after the year 1 BCE; there was no year 0.

Some other standards set by the Roman empire explain several abbreviations used in the text:
- c. is a Latin abbreviation for 'circa' and means 'roughly';
- i.e. is a Latin abbreviation for 'id est' and means 'that is';
- e.g. is a Latin abbreviation for 'exempli gratia' and means 'for the sake of example';
- ibid. is a Latin abbreviation for 'ibidem' and means 'at that same place';
- inf. is a Latin abbreviation for 'infra' and means '(see) below';
- op.cit. is a Latin abbreviation for 'opus citatum' and means 'the cited work';
- et al. is a Latin abbreviation for 'et alii' and means 'and others'.

By the way, 'idem' means 'the same' and 'passim' means 'here and there' or 'throughout'. Also terms like frequency, acceleration, velocity, mass, force, momentum, inertia, gravitation and temperature are Latin. In fact, there is a strong case to be made that the language of science has been Latin for over two thousand years. In Roman times it was Latin with Latin grammar, in modern times it switched to Latin vocabulary and French grammar, then for a short time to Latin with German grammar, after which it changed to Latin vocabulary and British/American grammar.

Many units of measurement also date from Roman times, as explained in the next appendix. Even the infatuation with Greek technical terms, as shown in coinages such as 'gyroscope', 'entropy' or 'proton', dates from Roman times.

Ref. 1130

Using such vocabulary is the best method to make language unintelligible to outsiders. First of all, it uses abbreviations, which is a shame. On top of this, the sentence uses people's names to characterize concepts, i.e. it uses *eponyms*. Originally, eponyms were intended as tributes to outstanding achievements. Today, when formulating new laws or variables has become nearly impossible, the spread of eponyms intelligible only to a steadily decreasing number of people simply reflects an increasingly ineffective drive to fame.

Eponyms are a lasting proof of the lack of imagination of scientists. Eponyms are avoided as much as possible in our walk; mathematical equations or entities are given *common* names wherever possible. People's names are then used as appositions to these names. For example, 'Newton's equation of motion' is never called 'Newton's equation', 'Einstein's field equations' is used instead of 'Einstein's equations', 'Heisenberg's equation of motion' in place of 'Heisenberg's equation'.

However, some exceptions are inevitable for certain terms within modern physics for which no real alternatives exist. The Boltzmann constant, the Planck scale, the Compton wavelength, the Casimir effect, Lie groups and the Virasoro algebra are examples. In compensation, the text makes sure that you can look up the definitions of these concepts using the index.

Bibliography

1119 For a clear overview of the various sign conventions in general relativity, see the front cover of Charles W. Misner, Kip S. Thorne & John A. Wheeler, *Gravitation*, Freeman, 1973. We use the gravitational sign conventions of Hans C. Ohanian & Remo Ruffini, *Gravitazione e spazio-tempo*, Zanichelli, 1997. Cited on page 1049.

1120 See for example the voice 'Mathematical notation' in the *Encyclopedia of Mathematics*, 10 volumes, Kluwer Academic Publishers, Dordrecht, 1988–1993. But first all, have a look at the informative and beautiful http://members.aol.com/jeff570/mathsym.html website. The main source for all these results is the classic and extensive research by Florian Cajori, *A History of Mathematical Notations*, 2 volumes, The Open Court Publishing Co., 1928–1929. Cited on page 1049.

1121 David R. Lide, editor, *CRC Handbook of Chemistry and Physics*, 78th edition, CRC Press, 1997. This classic reference work appears in a new edition every year. The full Hebrew alphabet is given on page 2-90. The list of abbreviations of physical quantities for use in formulae approved by iso, iupap and iupac can also be found there.

However, the iso 31 standard, which defines these abbreviations, costs around a thousand Euros, is not found on the internet, and therefore can safely be ignored, like any standard that is supposed to be used in teaching but is kept inaccessible to teachers. Cited on pages 1049 and 1054.

1122 Jan Tschichold, *Vormveranderingen van het &-teken*, Uitgeverij De Buitenkant, 1994. Cited on page 1051.

1123 Malcolm B. Parkes, *Pause and Effect: An Introduction to the History of Punctuation in the West*, University of California Press, 1993. Cited on page 1051.

1124 This is explained by Berthold Louis Ullman, *Ancient Writing and its Influence*, 1932. Cited on page 1051.

1125 Paul Lehmann, *Erforschung des Mittelalters – Ausgewählte Abhandlungen und Aufsätze*, Anton Hiersemann, Stuttgart, 1961, pp. 4–21. Cited on page 1051.

1126 Bernard Bischoff, *Paläographie des römischen Altertums und des abendländischen Mittelalters*, Erich Schmidt Verlag, 1979, pp. 215–219. Cited on page 1051.

1127 About the letters thorn and eth, have a look at the extensive report to be found on the website http://www.everytype.com/standards/wynnyogh/thorn.html. Cited on page 1052.

1128 Hans Jensen, *Die Schrift*, Berlin, 1969, translated into English as *Sign, Symbol and Script: an Account of Man's Efforts to Write*, Putnam's Sons, New York 1970. Cited on page 1052.

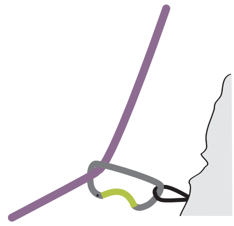
1129 See the mighty text by Peter T. Daniels & William Bright, *The World's Writing Systems*, Oxford University Press, 1996. Cited on page 1055.

1130 The connections between Greek roots and many French words – and thus many English ones – can be used to rapidly build up a vocabulary of ancient Greek without much study, as shown by the practical collection by J. Chaineux, *Quelques racines grecques*, Wetteren – De Meester, 1929. See also Donald M. Ayers, *English Words from Latin and Greek Elements*, University of Arizone Press, 1986. Cited on page 1057.

1131 To write well, read William Strunk & E.B. White, *The Elements of Style*, Macmillan, 1935, 1979, Wolf Schneider, *Deutsch für Kenner – Die neue Stilkunde*, Gruner und Jahr, 1987. No citations.

Appendix B

Units, Measurements and Constants



Measurements are comparisons. The standard used for comparison is called *unit*. Many different systems of units have been used throughout the world. All unit systems are standards and thus confer power to the organization in charge of them, as can be seen most clearly in the computer industry. In the past the same applied to measurement units. To avoid misuse by authoritarian institutions, to eliminate at the same time all problems with differing, changing and irreproducible standards, and – this is not a joke – to simplify tax collection, already in the eighteenth century scientists, politicians and economists have agreed on a set of units. It is called the *Système International d'Unités*, abbreviated *SI*, and is defined by an international treaty, the 'Convention du Mètre'. The units are maintained by an international organization, the 'Conférence Générale des Poids et Mesures', and its daughter organizations, the 'Commission Internationale des Poids et Mesures' and the 'Bureau International des Poids et Mesures' (BIPM), which all originated in the times just before the French revolution.

Ref. 1132

All SI units are built from seven *base units* whose official definitions, translated from French into English, are the following, together with the date of their formulation:

- 'The *second* is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.' (1967)*
- 'The *metre* is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second.' (1983)
- 'The *kilogram* is the unit of mass; it is equal to the mass of the international prototype of the kilogram.' (1901)*
- 'The *ampere* is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to $2 \cdot 10^{-7}$ newton per metre of length.' (1948)
- 'The *kelvin*, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water.' (1967)*
- 'The *mole* is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12.' (1971)*
- 'The *candela* is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency $540 \cdot 10^{12}$ hertz and has a radiant intensity in that direction of (1/683) watt per steradian.' (1979)*

---

* The international prototype of the kilogram is a platinum–iridium cylinder kept at the BIPM in Sèvres, in France. For more details on the levels of the caesium atom, consult a book on atomic physics. The Celsius scale of temperature θ is defined as: $\theta/°C = T/K - 273.15$; note the small difference with the number

Ref. 1133

Note that both time and length units are defined as certain properties of a standard example of motion, namely light. This is an additional example making the point that the observation of motion as the fundamental type of change is a *prerequisite* for the definition and construction of time and space. By the way, the proposal of using light was made already in 1827 by Jacques Babinet.*

From these basic units, all other units are defined by multiplication and division. In this way, all SI units have the following properties:

- SI units form a system with *state-of-the-art precision*; all units are defined in such a way that the precision of their definition is higher than the precision of commonly used measurements. Moreover, the precision of the definitions is regularly improved. The present relativeuncertainty of the definition of the second is around 10^{-14}, for the metre about 10^{-10}, for the kilogram about 10^{-9}, for the ampere 10^{-7}, for the mole less than 10^{-6}, for the kelvin 10^{-6} and for the candela 10^{-3}.
- SI units form an *absolute* system; all units are defined in such a way that they can be reproduced in every suitably equipped laboratory, independently, and with high precision. This avoids as much as possible any misuse by the standard setting organization. (At present, the kilogram, still defined with help of an artefact, is the last exception to this requirement; extensive research is under way to eliminate this artefact from the definition – an international race that will take a few more years.A definition can be based only on two ways: counting particles or fixing $\hbar$. The former can be achieved in crystals, the latter using any formula where $\hbar$ appears, such as the de Broglie wavelength, Josephson junctions, etc.)
- SI units form a *practical* system: base units are adapted to daily life quantities. Frequently used units have standard names and abbreviations. The complete list includes the seven base units, the derived, the supplementary and the admitted units.

The *derived* units with special names, in their official English spelling, i.e. without capital letters and accents, are:

appearing in the definition of the kelvin. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles. In its definition, it is understood that the carbon 12 atoms are unbound, at rest and in their ground state. The frequency of the light in the definition of the candela corresponds to 555.5 nm, i.e. green colour, and is the wavelength for which the eye is most sensitive.

* Jacques Babinet (1794–1874), French physicist who published important work in optics.

Name	Abbreviation	Name	Abbreviation
hertz	$Hz = 1/s$	newton	$N = kg\,m/s^2$
pascal	$Pa = N/m^2 = kg/m\,s^2$	joule	$J = Nm = kg\,m^2/s^2$
watt	$W = kg\,m^2/s^3$	coulomb	$C = As$
volt	$V = kg\,m^2/As^3$	farad	$F = As/V = A^2s^4/kg\,m^2$
ohm	$\Omega = V/A = kg\,m^2/A^2s^3$	siemens	$S = 1/\Omega$
weber	$Wb = Vs = kg\,m^2/As^2$	tesla	$T = Wb/m^2 = kg/As^2 = kg/Cs$
henry	$H = Vs/A = kg\,m^2/A^2s^2$	degree Celsius	°C (see definition of kelvin)
lumen	$lm = cd\,sr$	lux	$lx = lm/m^2 = cd\,sr/m^2$
becquerel	$Bq = 1/s$	gray	$Gy = J/kg = m^2/s^2$
sievert	$Sv = J/kg = m^2/s^2$	katal	$kat = mol/s$

We note that in all definitions of units, the kilogram only appears to the powers of 1, 0 and -1. The final explanation for this fact appeared only recently. Can you try to formulate the reason?

Challenge 1469 ny

The *supplementary* SI units are two: the unit for (plane) angle, defined as the ratio of arc length and radius, is the *radian* (rad). For solid angle, defined as the ratio of the subtended area and the square of the radius, the unit is the *steradian* (sr).

The *admitted* non-SI units are *minute*, *hour*, *day* (for time), *degree* $1° = \pi/180$ rad, *minute* $1' = \pi/10\,800$ rad, *second* $1'' = \pi/648\,000$ rad (for angles), *litre* and *tonne*. All other units are to be avoided.

All SI units are made more practical by the introduction of standard names and abbreviations for the powers of ten, the so-called *prefixes*:*

Power	Name		Power	Name		Power	Name		Power	Name	
10^1	deca	da	10^{-1}	deci	d	10^{18}	Exa	E	10^{-18}	atto	a
10^2	hecto	h	10^{-2}	centi	c	10^{21}	Zetta	Z	10^{-21}	zepto	z
10^3	kilo	k	10^{-3}	milli	m	10^{24}	Yotta	Y	10^{-24}	yocto	y
10^6	Mega	M	10^{-6}	micro	μ	unofficial:			Ref. 1134		
10^9	Giga	G	10^{-9}	nano	n	10^{27}	Xenta	X	10^{-27}	xenno	x
10^{12}	Tera	T	10^{-12}	pico	p	10^{30}	Wekta	W	10^{-30}	weko	w
10^{15}	Peta	P	10^{-15}	femto	f	10^{33}	Vendekta	V	10^{-33}	vendeko	v
						10^{36}	Udekta	U	10^{-36}	udeko	u

* Some of these names are invented (yocto to sound similar to Latin octo 'eight', zepto to sound similar to Latin septem, yotta and zetta to resemble them, exa and peta to sound like the Greek words of six and five, the unofficial ones to sound similar to the Greek words for nine, ten, eleven and twelve), some are from Danish/Norwegian (atto from atten 'eighteen', femto from femten 'fifteen'), some are from Latin (from mille 'thousand', from centum 'hundred', from decem 'ten', from nanus 'dwarf'), some are from Italian (from piccolo 'small'), some are Greek (micro is from μικρός 'small', deca/deka from δέκα 'ten', hecto from ἑκατόν 'hundred', kilo from χίλιοι 'thousand', mega from μέγας 'large', giga from γίγας 'giant', tera from τέρας 'monster').

Translate: I was caught in such a traffic jam that I needed a microcentury for a picoparsec and that my car's fuel consumption was two tenths of a square millimetre.

Challenge 1470 e

▪ SI units form a *complete* system; they cover in a systematic way the complete set of observables of physics. Moreover, they fix the units of measurements for physics and for all other sciences as well.

▪ SI units form a *universal* system; they can be used in trade, in industry, in commerce, at home, in education and in research. They could even be used by other civilizations, if they existed.

▪ SI units form a *coherent* system; the product or quotient of two SI units is also a SI unit. This means that in principle, the same abbreviation 'SI' could be used for every SI unit.

The SI units are not the only possible set that fulfils all these requirements, but they form the only existing system doing so.*

We remind that since every measurement is a comparison with a standard, any measurement requires matter to realize the standard (yes, even for the speed standard) and radiation to achieve the comparison. Our concept of measurement thus assumes that matter and radiation exist and can be clearly separated from each other. Page 1019

Planck's natural units

Since the exact form of many equations depends on the used system of units, theoretical physicists often use unit systems optimized for producing simple equations. In microscopic physics, the system of *Planck's natural units* is frequently used. They are automatically introduced by setting $c = 1$, $\hbar = 1$, $G = 1$, $k = 1$, $\varepsilon_0 = 1/4\pi$ and $\mu_0 = 4\pi$ in equations written in SI units. Planck units are thus defined from combinations of fundamental constants; those corresponding to the fundamental SI units are given in the table.** The table is also useful for converting equations written in natural units back to SI units; in this case, every quantity X is substituted by X/X_{Pl}. Challenge 1471 e

Table 79 Planck's (uncorrected) natural units

Name	Definition				Value
Basic units					
the Planck length	l_{Pl}	=	$\sqrt{\hbar G/c^3}$	=	$1.616\,0(12) \cdot 10^{-35}$ m
the Planck time	t_{Pl}	=	$\sqrt{\hbar G/c^5}$	=	$5.390\,6(40) \cdot 10^{-44}$ s
the Planck mass	m_{Pl}	=	$\sqrt{\hbar c/G}$	=	$21.767(16)$ µg

* Most non-SI units still in use in the world are of Roman origin: the mile comes from 'milia passum' (used to be one thousand (double) strides of about 1480 mm each; today a nautical mile, after having been defined as minute of arc, is exactly 1852 m), inch comes from 'uncia/onzia' (a twelfth – now of a foot); pound (from pondere 'to weigh') is used as a translation of 'libra' – balance – which is the origin of its abbreviation *lb*; even the habit of counting in dozens instead of tens is Roman in origin. These and all other similarly funny units – like the system in which all units start with 'f' and which uses furlong/fortnight as unit for velocity – are now officially defined as multiples of SI units.

** The natural units x_{Pl} given here are those commonly used today, i.e. those defined using the constant $\hbar$, and not, as Planck originally did, by using the constant $h = 2\pi\hbar$. A similar, additional freedom of choice arises for the electromagnetic units, which can be defined with other factors than $4\pi\varepsilon_0$ in the expressions; for example, using $4\pi\varepsilon_0\alpha$, with the *fine structure constant* α, gives $q_{\text{Pl}} = e$. For the explanation of the numbers between brackets, the standard deviations, see page 1069.

Name	Definition			Value	
the Planck current	I_{Pl}	=	$\sqrt{4\pi\varepsilon_0 c^6/G}$	=	$3.4793(22)\cdot 10^{25}$ A
the Planck temperature	T_{Pl}	=	$\sqrt{\hbar c^5/Gk^2}$	=	$1.4171(91)\cdot 10^{32}$ K
Trivial units					
the Planck velocity	v_{Pl}	=	c	=	0.3 Gm/s
the Planck angular momentum	L_{Pl}	=	$\hbar$	=	$1.1\cdot 10^{-34}$ Js
the Planck action	S_{aPl}	=	$\hbar$	=	$1.1\cdot 10^{-34}$ Js
the Planck entropy	S_{ePl}	=	k	=	13.8 yJ/K
Composed units					
the Planck mass density	ρ_{Pl}	=	$c^5/G^2\hbar$	=	$5.2\cdot 10^{96}$ kg/m^3
the Planck energy	E_{Pl}	=	$\sqrt{\hbar c^5/G}$	=	2.0 GJ = $1.2\cdot 10^{28}$ eV
the Planck momentum	p_{Pl}	=	$\sqrt{\hbar c^3/G}$	=	6.5 Ns
the Planck force	F_{Pl}	=	c^4/G	=	$1.2\cdot 10^{44}$ N
the Planck power	P_{Pl}	=	c^5/G	=	$3.6\cdot 10^{52}$ W
the Planck acceleration	a_{Pl}	=	$\sqrt{c^7/\hbar G}$	=	$5.6\cdot 10^{51}$ m/s^2
the Planck frequency	f_{Pl}	=	$\sqrt{c^5/\hbar G}$	=	$1.9\cdot 10^{43}$ Hz
the Planck electric charge	q_{Pl}	=	$\sqrt{4\pi\varepsilon_0 c\hbar}$	=	1.9 aC = 11.7 e
the Planck voltage	U_{Pl}	=	$\sqrt{c^4/4\pi\varepsilon_0 G}$	=	$1.0\cdot 10^{27}$ V
the Planck resistance	R_{Pl}	=	$1/4\pi\varepsilon_0 c$	=	30.0 Ω
the Planck capacitance	C_{Pl}	=	$4\pi\varepsilon_0\sqrt{\hbar G/c^3}$	=	$1.8\cdot 10^{-45}$ F
the Planck inductance	L_{Pl}	=	$(1/4\pi\varepsilon_0)\sqrt{\hbar G/c^7}$	=	$1.6\cdot 10^{-42}$ H
the Planck electric field	E_{Pl}	=	$\sqrt{c^7/4\pi\varepsilon_0\hbar G^2}$	=	$6.5\cdot 10^{61}$ V/m
the Planck magnetic flux density	B_{Pl}	=	$\sqrt{c^5/4\pi\varepsilon_0\hbar G^2}$	=	$2.2\cdot 10^{53}$ T

The natural units are important for another reason: whenever a quantity is sloppily called 'infinitely small (or large)', the correct expression is 'small (or large) as the corresponding corrected Planck unit'. As explained throughout the text and especially in the third part, this substitution is possible because almost all Planck units provide, within a correction factor of order 1, the extreme value for the corresponding observable. Unfortunately, these correction factors have not entered the mainstream yet. The exact extremal value for each observable in nature is obtained when G is substituted by $4G$, $\hbar$ by $\hbar/2$, k by $k/2$ and $4\pi\varepsilon_0$ by $8\pi\varepsilon_0\alpha$ in all Planck quantities. These extremal values, or *corrected Planck units*, are the *true natural units*. Exceeding the extremal values is possible only for some extensive quantities. (Can you find out which ones?)

Page 993

Challenge 1472 n

Other unit systems

In high energy physics another unit system is also common. A central aim of research is the calculation of the strength of all interactions; therefore it is not practical to set the gravitational constant G to unity, as Planck units do. For this reason, high energy physicists often only set $c = \hbar = k = 1$ and $\mu_0 = 1/\varepsilon_0 = 4\pi$,* leaving only the gravitational constant G in the equations. In this system, only one fundamental unit exists, but its choice is still free.

Often a standard length is chosen as fundamental unit, length being the archetype of a measured quantity. The most important physical observables are then related by

$$\begin{aligned}
1/[l^2] &= [E]^2 = [F] = [B] = [E_{\text{electric}}] , \\
1/[l] &= [E] = [m] = [p] = [a] = [I] = [U] = [T] , \\
1 &= [v] = [q] = [e] = [R] = [S_{\text{action}}] = [S_{\text{entropy}}] = \hbar = c = k = [\alpha] , \\
[l] &= 1/[E] = [t] = [C] = [L] \quad \text{and} \\
[l]^2 &= 1/[E]^2 = [G] = [P]
\end{aligned} \tag{788}$$

with the usual convention to write $[x]$ for the unit of quantity x. Using the same unit for time, capacitance and inductance is not to everybody's taste, however, and therefore electricians do not use this system.*

In many situations, in order to get an impression of the energies needed to observe the effect under study, a standard energy is chosen as fundamental unit. In particle physics the common energy unit is the *electron Volt* (eV), defined as the kinetic energy acquired by an electron when accelerated by an electrical potential difference of 1 Volt ('proton Volt' would be a better name). Therefore one has $1\,\text{eV}=1.6 \cdot 10^{-19}\,\text{J}$, or roughly

$$1\,\text{eV} \approx \tfrac{1}{6}\ \text{aJ} \tag{789}$$

which is easily remembered. The simplification $c = \hbar = 1$ yields $G = 6.9 \cdot 10^{-57}\,\text{eV}^{-2}$ and allows to use the unit eV also for mass, momentum, temperature, frequency, time and length, with the respective correspondences $1\,\text{eV} \equiv 1.8 \cdot 10^{-36}\,\text{kg} \equiv 5.4 \cdot 10^{-28}\,\text{Ns} \equiv 242\,\text{THz} \equiv 11.6\,\text{kK}$ and $1\,\text{eV}^{-1} \equiv 4.1\,\text{fs} \equiv 1.2\,\mu\text{m}$. Challenge 1474 e

To get some feeling for the unit eV, the following relations are useful. Room temperature, usually taken as 20°C or 293 K, corresponds to a kinetic energy per particle of 0.025 eV or 4.0 zJ. The highest particle energy measured so far is a cosmic ray of energy of $3 \cdot 10^{20}$ eV or 48 J. Down here on the Earth, an accelerator with an energy of about Ref. 1136

* Other definitions for the proportionality constants in electrodynamics lead to the Gaussian unit system often used in theoretical calculations, the Heaviside–Lorentz unit system, the electrostatic unit system, and the electromagnetic unit system, among others. For more details, see the standard text by John David Jackson, *Classical Electrodynamics*, third edition, Wiley, 1998.

* In the list, l is length, E energy, F force, E_{electric} the electric and b the magnetic field, m mass, p momentum, a acceleration, I electric current, U voltage, T temperature, v speed, q charge, R resistance, P power, G the gravitational constant.

The web page http://www.chemie.fu-berlin.de/chemistry/general/units-en.html allows to convert various units into each other.

Researchers in general relativity often use another system, in which the *Schwarzschild radius* $r_S = 2Gm/c^2$ is used to measure masses, by setting $c = G = 1$. In this case, in opposition to above, mass and length have the *same* dimension, and $\hbar$ has the dimension of an area.

Ref. 1135 Already in the nineteenth century, George Stoney, had proposed to use as length, time and mass units the expressions $l_S = \sqrt{Ge^2/(c^4 4\pi\varepsilon_0)} = 1.4 \cdot 10^{-36}$ m, $t_S = \sqrt{Ge^2/(c^6 4\pi\varepsilon_0)} = 4.6 \cdot 10^{-45}$ s and $m_S = \sqrt{e^2/(G 4\pi\varepsilon_0)} = 1.9\,\mu$g. How are these units related to the Planck units? Challenge 1473 n

105 GeV or 17 nJ for electrons and antielectrons has been built, and one with an energy of 10 TeV or 1.6 μJ for protons will be finished soon.Both are owned by CERN in Geneva and have a circumference of 27 km.

The lowest temperature measured up to nowis 280 pK, in a system of Rhodium nuclei inside a special cooling system. The interior of that cryostat possibly is the coolest point in the whole universe. At the same time, the kinetic energy per particle corresponding to that temperature is also the smallest ever measured; it corresponds to 24 feV or 3.8 vJ=$3.8 \cdot 10^{-33}$ J. For isolated particles, the record seems to be for neutrons: kinetic energies as low as 10^{-7} eV have been achieved, corresponding to de Broglie wavelengths of 60 nm.

Ref. 1137

Curiosities

Here are a few facts making the concept of unit more vivid.

▪ A gray is the amount of radioactivity that deposes 1 J on 1 kg of matter. A sievert is a unit adjusted to human scale, where the different types of human tissues are weighted with a factor describing the effectiveness of radiation deposition. Four to five sievert are a lethal dose to humans. In comparison, the natural radioactivity present inside human bodies leads to a dose of 0.2 mSv per year. An average X-ray image implies an irradiation of 1 mSv; a CAT scan 8 mSv.

Ref. 1138

▪ Are you confused by the candela? The definition simply says that 683 cd = 683 lm/sr correspond to 1 W/sr. The candela is thus a unit for light power per angle, except that it is corrected for the eye's sensitivity: the candela measures only *visible* power per angle. Similarly, 683 lm = 683 cd · sr correspond to 1 W, i.e. both the lumen and the watt measure power, or energy flux, except that the lumen measures only the *visible* part of the power. In English quantity names, the difference is expressed by substituting 'radiant' by 'luminous'; e.g. the Watt measures *radiant* flux, whereas the lumen measure *luminous* flux.

The factor 683 is historical. A usual candle indeed emits a luminous intensity of about a candela. Therefore, at night, a candle can be seen up to a distance of one or two dozen kilometres. A 100 W incandescent light bulb produces 1700 lm and the brightest light emitting diodes about 5 lm. Cinema projectors produce around 2 Mlm, and the brightest flashes, like lightning, 100 Mlm.

Challenge 1475 e

The *irradiance* of sunlight is about 1300 W/m^2 on a sunny day; on the other hand, the *illuminance* is only 120 klm/m^2 = 120 klux or 170 W/m^2. (A cloud-covered summer day or a clear winter day produces about 10 klux.) The numbers show that most energy radiated from the Sun to the Earth is outside the visible spectrum.

On a glacier, near the sea shore, on the top of mountains, or under particular weather condition the brightness can reach 150 klux. The brightest lamps, those used during surgical operations, produce 120 klux. Humans need about 30 lux for comfortable reading. Museums are often dark because water paintings are destroyed by more than 100 lux, oil paintings by more than 200 lux. The full Moon produces 0.1 lux, and a dark moonless night has about 1 mlux. the eyes lose their ability to distinguish colours somewhere between 0.1 lux and 0.01 lux; the eye starts to work at 1 nlux. Devices to see in the dark start to work at 1 μlux By the way, the human body itself *shines* with about 1 plux, a value too small to be detected with the eye, but easily measured with apparatus. The origin of this emission is still a topic of research.

Ref. 1139

▪ The highest achieved light intensities are in excess of 10^{18} W/m^2, more than 15 orders of magnitude higher than the intensity of sunlight. They are produced by tight focusing of pulsed lasers. The electric field in such light pulses is of the same order of the field inside atoms; such a beam therefore ionizes all matter it encounters. Ref. 1140

▪ The luminous density is often used by light technicians. Its unit is 1 cd/m^2, unofficially called 1 Nit and abbreviated 1 nt. Eyes see purely with rods from 0.1 µcd/m^2 to 1 mcd/m^2, see purely with cones above 5 cd/m^2, see best between 100 to 50 000 cd/m^2, and get fully overloaded above 10 Mcd/m^2, a total range of 15 orders of magnitude.

▪ The Planck length is roughly the de Broglie wavelength $\lambda_B = h/mv$ of a man walking comfortably ($m = 80$ kg, $v = 0.5$ m/s); this motion is therefore aptly called the 'Planck stroll.' Ref. 1141

▪ The Planck mass is equal to the mass of about 10^{19} protons. This is roughly the mass of a human embryo at about ten days of age.

▪ The second does not correspond to 1/86 400th of the day any more (it did so in the year 1900); the Earth now takes about 86 400.002 s for a rotation, so that regularly the *International Earth Rotation Service* introduces a leap second to ensure that the Sun is at the highest point in the sky at 12.00 o'clock sharp.* The time so defined is called *Universal Time Coordinate*. The velocity of rotation of the Earth also changes irregularly from day to day due to the weather; the average rotation speed even changes from winter to summer due to the change in polar ice caps and in addition that average decreases over time, due to the friction produced by the tides. The rate of insertion of leap seconds is therefore faster than every 500 days, and not completely constant in time.

▪ The most precisely measured quantities in nature are the frequencies of certain millisecond pulsars,** the frequency of certain narrow atomic transitions and the Rydberg constant of *atomic* hydrogen, which can all be measured as exactly as the second is defined. At present, this gives about 14 digits of precision.

▪ The most precise clock ever built, using microwaves, had a stability of 10^{-16} during a running time of 500 s. For longer time periods, the record in 1997 was about 10^{-15}; but the area of 10^{-17} seems within technological reach. The precision of clocks is limited for short measuring times by noise and for long measuring times by drifts, i.e. by systematic effects. The region of highest stability depends on the clock type and usually lies between 1 ms for optical clocks and 5000 s for masers. Pulsars are the only clock for which this region is not known yet; it lies at more than 20 years, which is the time elapsed since their discovery. Ref. 1142 Ref. 1143

▪ The shortest times measured are the life times of certain 'elementary' particles; in particular, the D meson was measured to live less than 10^{-22} s. Such times are measured in a bubble chamber, where the track is photographed. Can you estimate how long the track is? (Watch out – if your result cannot be observed with an optical microscope, you made a mistake in your calculation). Ref. 1144 Challenge 1476 n

▪ The longest measured times are the lifetimes of certain radioisotopes, over 10^{15} years, and the lower limit of certain proton decays, over 10^{32} years. These times are thus much

* Their website at http://hpiers.obspm.fr gives more information on the details of these insertions, as does http://maia.usno.navy.mil, one of the few useful military websites. See also http://www.bipm.fr, the site of the BIPM.

** An overview of this fascinating work is given by J.H. TAYLOR, Pulsar timing and relativistic gravity, *Philosophical Transactions of the Royal Society, London A* **341**, pp. 117–134, 1992.

Ref. 1145 larger than the age of the universe, estimated to be fourteen thousand million years.

▪ The least precisely measured fundamental quantities are the gravitational constant G and the strong coupling constant α_s. Other, even less precisely known quantities, are Page 1071 the age of the universe and its density (see the astrophysical table below).

▪ The precision of mass measurements of solids is limited by such simple effects as the adsorption of water on the weight. Can you estimate what a monolayer of water does on Challenge 1477 n a weight of 1 kg?

▪ Variations of quantities are often much easier to measure than their values. For example, in gravitational wave detectors, the sensitivity achieved in 1992 was $\Delta l/l = 3 \cdot 10^{-19}$ Ref. 1146 for lengths of the order of 1 m. In other words, for a block of about a cubic metre of metal it is possible to measure length changes about 3000 times smaller than a proton radius. These set-ups are now being superseded by ring interferometers. Ring interferometers measuring frequency differences of 10^{-21} have already been built; they are still being im- Ref. 1147 proved towards higher values.

▪ The swedish astronomer Anders Celsius (1701–1744) originally set the freezing point at 100 degrees and the boiling point of water at 0 degrees. Then the numbers were switched Ref. 1148 to get today's scale. However, this is not the whole story. With the official definition of the Kelvin and the degree Celsius, at the standard pressure of 1013.25 Pa, water boils at Challenge 1478 n 99.974 °C. Can you explain why it is not 100 °C any more?

▪ In the old days, thermal energy used to be measured in the unit *calorie*, written as cal. It is the energy needed to heat 1 g of water by 1 K. To confuse matters, 1 kcal was often abbreviated 1 Cal. (One also spoke of a large and a small calorie.) One has 1 kcal = 4.1868 kJ.

▪ SI units are adapted to humans: the values of heartbeat, human size, human weight, human temperature and human substance are mostly near the unit value. Si units this (roughly) realize the saying by Protagoras 25 centuries ago: 'Man is the measure of all things.'

▪ The table of SI prefixes covers seventy-two measurement decades. How many additional prefixes will be needed? Even an extended list will include only a small part of the infinite range of possibilities. Will the Conférence Générale des Poids et Mesures have to Challenge 1479 n go on and on, defining an infinite number of SI prefixes?

▪ It is well-known that the French philosopher Voltaire, after meeting Newton, publicized the now famous story that the connection between the fall of objects and the motion of the Moon was discovered by Newton when he saw an apple falling from a tree. More than a century later, just before the French revolution, a committee of scientists decided to take as unit of force precisely the force exerted by gravity on a *standard apple*, and to name it after the English scientist. After extensive study, it was found that the mass of the standard apple was 101.9716 g; its weight was called 1 newton. Since then, in the museum in Sèvres near Paris, visitors can admire the standard metre, the standard kilogram and the standard apple.*

* To be clear, this is invented; no standard apple exists. In contrast to the apple story it is *not* a joke however, that owners of several apple trees in Britain and in the US claim descent, by rerooting, from the original tree Ref. 1149 under which Newton had his insight. DNA tests have even been performed to decide if all these derive from the same tree, with the result that the tree at MIT, in contrast to the British ones, is a fake – of course.

Precision and accuracy of measurements

As explained on page 240, *precision* measures how well a result is reproduced when the measurement is repeated; *accuracy* is the degree to which a measurement corresponds to the actual value. Lack of precision is due to accidental or *random errors*; they are best measured by the *standard deviation*, usually abbreviated σ; it is defined through

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (790)$$

where $\bar{x}$ is the average of the measurements x_i. (Can you imagine why $n-1$ is used in the formula instead of n?) By the way, for a Gaussian distribution, 2.35 σ is the full width at half maximum.

Challenge 1480 n

Lack of accuracy is due to *systematic errors*; usually they can only be estimated. This estimate is often added to the random errors to produce a *total* experimental error, sometimes also called *total uncertainty*.

Ref. 1150

The following tables give the values of the most important physical constants and particle properties in SI units and in a few other common units, as published in the standard references. The values are the world average of the best measurements up to the present. As usual, experimental errors, including both random and estimated systematic errors, are expressed by giving the one standard deviation uncertainty in the last digits; e.g. 0.31(6) means – roughly speaking – 0.31 ± 0.06. In fact, behind each of the numbers in the following tables there is a long story which is worth telling, but for which there is not enough room here.*

Ref. 1151

What are the limits to accuracy and precision? There is no way, even in principle, to measure a quantity x to a *precision* higher than about 61 digits, because $\Delta x/x > l_{\mathrm{Pl}}/d_{\mathrm{horizon}} = 10^{-61}$. (Is this correct for force or for volume?) In the third part of our text, studies of clocks and meter bars will further reduce this theoretical limit.

Challenge 1481 ny
Page 955

But it is not difficult to deduce more stringent practical limits. No reasonable machine can measure quantities with a higher precision than measuring the diameter of the Earth within the smallest length ever measured, about 10^{-19} m; that makes about 26 digits. Using a more realistic limit of a 1000 m sized machine implies a limit of 22 digits. If, as predicted above, time measurements really achieve 17 digits of precision, then they are nearing the practical limit, because apart from size, there is an additional practical restriction: cost. Indeed, an additional digit in measurement precision often means an additional digit in equipment cost.

Basic physical constants

Ref. 1151

In principle, all experimental measurements of matter properties can be calculated with quantum theory. For example, colour, density or elastic properties, can be predicted using

* Some of the stories can be found in the text by N.W. Wise, *The Values of Precision*, Princeton University Press, 1994. The field of high precision measurements, from which the results on these pages stem, is a world on its own. A beautiful introduction to it is *Near Zero: Frontiers of Physics*, edited by J.D. Fairbanks, B.S. Deaver, C.W. Everitt & P.F. Michaelson, Freeman, 1988.

the values of the following constants, using the equations of the standard model of high energy physics. Page ??

Table 80 Basic physical constants

Quantity	Name	Value in SI units	Uncert.
vacuum number of space-time dimensions		$3+1$	down to 10^{-19} m, up to 10^{26} m
vacuum speed of light[a]	c	299 792 458 m/s	0
vacuum permeability[a]	μ_0	$4\pi \cdot 10^{-7}$ H/m	0
		= 1.256 637 061 435 917 295 ... μH/m	
vacuum permittivity[a]	$\varepsilon_0 = 1/\mu_0 c^2$	8.854 187 817 620 ... pF/m	0
original Planck constant	h	$6.626\,068\,76(52) \cdot 10^{-34}$ Js	$7.8 \cdot 10^{-8}$
reduced Planck constant	$\hbar$	$1.054\,571\,596(82) \cdot 10^{-34}$ Js	$7.8 \cdot 10^{-8}$
positron charge	e	$0.160\,217\,646\,2(63)$ aC	$3.9 \cdot 10^{-8}$
Boltzmann constant	k	$1.380\,650\,3(24) \cdot 10^{-23}$ J/K	$1.7 \cdot 10^{-6}$
gravitational constant	G	$6.673(10) \cdot 10^{-11}$ Nm²/kg²	$1.5 \cdot 10^{-3}$
gravitational coupling constant	$\kappa = 8\pi G/c^4$	$2.076(3) \cdot 10^{-43}$ s²/kg m	$1.5 \cdot 10^{-3}$
fine structure constant,[b]	$\alpha = \frac{e^2}{4\pi\varepsilon_0 \hbar c}$	$1/137.035\,999\,76(50)$	$3.7 \cdot 10^{-9}$
e.m. coupling constant	$= g_{em}(m_e^2 c^2)$	$= 0.007\,297\,352\,533(27)$	$3.7 \cdot 10^{-9}$
Fermi coupling constant,[b]	$G_F/(\hbar c)^3$	$1.166\,39(1) \cdot 10^{-5}$ GeV^{-2}	$8.6 \cdot 10^{-6}$
weak coupling constant	$\alpha_w(M_Z) = g_w^2/4\pi$	$1/30.1(3)$	
weak mixing angle	$\sin^2 \theta_W(\overline{MS})$	$0.231\,24(24)$	$1.0 \cdot 10^{-3}$
weak mixing angle	$\sin^2 \theta_W$ (on shell)	$0.2224(19)$	$8.7 \cdot 10^{-3}$
	$= 1 - (m_W/m_Z)^2$		
strong coupling constant[b]	$\alpha_s(M_Z) = g^2 s/4\pi$	$0.118(3)$	$25 \cdot 10^{-3}$

a. Defining constant.
b. All coupling constants depend on the four-momentum transfer, as explained in the section on renormalization. *Fine structure constant* is the traditional name for the electromagnetic coupling constant g_{em} in the case of a four momentum transfer of $Q^2 = m_e^2 c^2$, which is the smallest one possible. At higher momentum transfers it has larger values, e.g. $g_{em}(Q^2 = M_W^2 c^2) \approx 1/128$. The strong coupling constant has higher values at lower momentum transfers; e.g. one has $\alpha_s(34\,\text{GeV}) = 0.14(2)$. Page 802

Why do all these constants have the values they have? The answer depends on the constant. For any constant having a unit, such as the quantum of action $\hbar$, the numerical value has only historical meaning. It is $1.054 \cdot 10^{-34}$ Js because of the SI definition of the joule and the second. The question why the value of a constant with units is not larger or smaller always requires one to understand the origin of some dimensionless number giving the ratio between the constant and the corresponding natural unit. Understanding the size of atoms, people, trees and stars, the duration of molecular and atomic processes, or the mass of nuclei and mountains implies to understand the ratio between these values and the corresponding natural units. The key to understanding nature is thus the understanding of all ratios, thus of all dimensionless constants. The story of the most important ratios is told in the main text.

The basic constants yield the following useful high-precision observations.

Table 81 Derived physical constants

Quantity	Name	Value in SI units	Uncert.
Vacuum wave resistance	$Z_o = \sqrt{\mu_0/\varepsilon_0}$	376.730 313 461 77... Ω	0
Avogadro's number	N_A	$6.022\,141\,99(47) \cdot 10^{23}$	$7.9 \cdot 10^{-8}$
Rydberg constant [a]	$R_\infty = m_e c\alpha^2/2h$	$10\,973\,731.568\,549(83)\ \mathrm{m}^{-1}$	$7.6 \cdot 10^{-12}$
conductance quantum	$G_0 = 2e^2/h$	77.480 916 96(28) μS	$3.7 \cdot 10^{-9}$
mag. flux quantum	$\varphi_0 = h/2e$	2.067 833 636(81) pWb	$3.9 \cdot 10^{-8}$
Josephson freq. ratio	$2e/h$	483.597 898(19) THz/V	$3.9 \cdot 10^{-8}$
von Klitzing constant	$h/e^2 = \mu_0 c/2\alpha$	25 812.807 572(95) Ω	$3.7 \cdot 10^{-9}$
Bohr magneton	$\mu_B = e\hbar/2m_e$	9.274 008 99(37) yJ/T	$4.0 \cdot 10^{-8}$
cyclotron frequency of the electron	$f_c/B = e/2\pi m_e$	27.992 4925(11) GHz/T	$4.0 \cdot 10^{-8}$
classical electron radius	$r_e = e^2/4\pi\varepsilon_0 m_e c^2$	2.817 940 285(31) fm	$1.1 \cdot 10^{-8}$
Compton wavelength	$\lambda_c = h/m_e c$	2.426 310 215(18) pm	$7.3 \cdot 10^{-9}$
of the electron	$\bar{\lambda}_c = \hbar/m_e c = r_e/\alpha$	0.386 159 264 2(28) pm	$7.3 \cdot 10^{-9}$
Bohr radius [a]	$a_\infty = r_e/\alpha^2$	52.917 720 83(19) pm	$3.7 \cdot 10^{-9}$
nuclear magneton	$\mu_N = e\hbar/2m_p$	$5.050\,783\,17(20) \cdot 10^{-27}$ J/T	$4.0 \cdot 10^{-8}$
proton electron mass ratio	m_p/m_e	1 836.152 667 5(39)	$2.1 \cdot 10^{-9}$
Stefan–Boltzmann constant	$\sigma = \pi^2 k^4/60\hbar^3 c^2$	$56.704\,00(40)\ \mathrm{nW/m^2K^4}$	$7.0 \cdot 10^{-6}$
Wien displacement law constant	$b = \lambda_{max} T$	2.897 768 6(51) mmK	$1.7 \cdot 10^{-6}$
bits to entropy conv. const.		10^{23} bit = 0.956 994 5(17) J/K	$1.7 \cdot 10^{-6}$
TNT energy content		3.7 to 4.0 MJ/kg=$4 \cdot 10^3\ \mathrm{m^2/s^2}$	$4 \cdot 10^{-2}$

a. For infinite mass of the nucleus.

Some properties of nature at large are listed in the following table. (If you want a challenge, can you state whether any property of the universe itself is listed?)

Challenge 1482 n

Table 82 Astrophysical constants

Quantity	Name	Value
gravitational constant	G	$6.672\,59(85) \cdot 10^{-11}\ \mathrm{m^3/kg\,s^2}$
cosmological constant	Λ	*c.* $1 \cdot 10^{-52}\ \mathrm{m}^{-2}$
tropical year 1900 [a]	a	31 556 925.974 7 s
tropical year 1994	a	31 556 925.2 s
mean sidereal day	d	$23^h 56' 4.090\,53''$
light year	al	9.460 528 173 ... Pm
astronomical unit [b]	AU	149 597 870.691(30) km
parsec	pc	30.856 775 806 Pm = 3.261 634 al

QUANTITY	NAME	VALUE
age of the universe [c]	t_0	$> 3.5(4) \cdot 10^{17}$ s or $> 11.5(1.5) \cdot 10^9$ a
(from matter, via galaxies and stars, using quantum theory)		
age of the universe [c]	t_0	$4.32(7) \cdot 10^{17}$ s $= 13.7(0.2) \cdot 10^9$ a
(from space-time, via expansion, using general relativity)		
Hubble parameter [c]	H_0	$2.3(2) \cdot 10^{-18}\,\mathrm{s}^{-1} = 0.73(4) \cdot 10^{-10}\,\mathrm{a}^{-1}$
	$H_0 = h_0 \cdot 100\,\mathrm{km/sMpc} = h_0 \cdot 1.0227 \cdot 10^{-10}\,\mathrm{a}^{-1}$	
reduced Hubble par. [c]	h_0	0.71(4)
deceleration parameter	$q_0 = -(\ddot{a}/a)_0/H_0^2$	-0.66(10)
universe's horizon's dist. [c]	$d_0 = 3ct_0$	$40.0(6) \cdot 10^{26}$ m=13.0(2) Gpc
universe's topology		unknown
number of space dimensions		3, up to 10^{26} m
critical density	$\rho_c = 3H_0^2/8\pi G$	$h_0^2 \cdot 1.878\,82(24) \cdot 10^{-26}\,\mathrm{kg/m^3}$
of the universe		$= 0.95(12) \cdot 10^{-26}\,\mathrm{kg/m^3}$
(total) density parameter [c]	$\Omega_0 = \rho_0/\rho_c$	1.02(2)
baryon density parameter [c]	$\Omega_{\mathrm{Bo}} = \rho_{\mathrm{Bo}}/\rho_c$	0.044(4)
cold dark matter density parameter [c]	$\Omega_{\mathrm{CDMo}} = \rho_{\mathrm{CDMo}}/\rho_c$	0.23(4)
neutrino density parameter [c]	$\Omega_{\nu\mathrm{o}} = \rho_{\nu\mathrm{o}}/\rho_c$	0.001 to 0.05
dark energy density parameter [c]	$\Omega_{\mathrm{Xo}} = \rho_{\mathrm{Xo}}/\rho_c$	0.73(4)
dark energy state equation	$w = p_X/\rho_X$	-1.0(2)
baryonmass	m_b	$1.67 \cdot 10^{-27}$ kg
baryon number density		$0.25(1)\,/\mathrm{m^3}$
luminous matter density		$3.8(2) \cdot 10^{-28}\,\mathrm{kg/m^3}$
stars in the universe	n_s	$10^{22\pm1}$
baryons in the universe	n_b	$10^{81\pm1}$
microwave background temperature [d]	T_0	2.725(1) K
photons in the universe	n_γ	10^{89}
photon energy density	$\rho_\gamma = \pi^2 k^4/15T_0^4$	$4.6 \cdot 10^{-31}\,\mathrm{kg/m^3}$
photon number density		$410.89\,/\mathrm{cm^3}$ or $400\,/\mathrm{cm^3}(T_0/2.7\,\mathrm{K})^3$
density perturbation amplitude	$\sqrt{S}$	$5.6(1.5) \cdot 10^{-6}$
gravity wave amplitude	$\sqrt{T}$	$< 0.71\sqrt{S}$
mass fluctuations on 8 Mpc	σ_8	0.84(4)
scalar index	n	0.93(3)
running of scalar index	$\mathrm{d}n/\mathrm{d}\ln k$	-0.03(2)
Planck length	$l_{\mathrm{Pl}} = \sqrt{\hbar G/c^3}$	$1.62 \cdot 10^{-35}$ m
Planck time	$t_{\mathrm{Pl}} = \sqrt{\hbar G/c^5}$	$5.39 \cdot 10^{-44}$ s
Planck mass	$m_{\mathrm{Pl}} = \sqrt{\hbar c/G}$	21.8 µg
instants in history [c]	t_0/t_{Pl}	$8.7(2.8) \cdot 10^{60}$
space-time points	$N_0 = (R_0/l_{\mathrm{Pl}})^3 \cdot$	$10^{244\pm1}$
inside the horizon [c]	(t_0/t_{Pl})	

Quantity	Name	Value
mass inside horizon	M	$10^{54\pm1}$ kg

a. Defining constant, from vernal equinox to vernal equinox; it was once used to define the second. (Remember: π seconds is a nanocentury.) The value for 1990 is about 0.7 s less, corresponding to a slowdown of roughly −0.2 ms/a. (Watch out: why?) There is even an empirical formula available for the change of the length of the year over time.

Challenge 1483 n
Ref. 1152

b. Average distance Earth–Sun. The truly amazing precision of 30 m results from time averages of signals sent from Viking orbiters and Mars landers taken over a period of over twenty years.

c. The index o indicates present day values.

d. The radiation originated when the universe was between 10^5 to 10^6 years old and had a temperature of about 3000 K; the fluctuations ΔT_0 which lead to galaxy formation are today of the size of $16 \pm 4\,\mu K = 6(2) \cdot 10^{-6}\ T_o$.

Page 415

Attention: in the third part of this text it is shown that many constants in Table 82 are *not* physically sensible quantities. They have to be taken with lots of grains of salt. The more specific constants given in the following table are all sensible though.

Table 83 Astronomical constants

Quantity	Name	Value
Earth's mass	M	$5.972\,23(8) \cdot 10^{24}$ kg
Earth's gravitational length	$l = 2GM/c^2$	8.870(1) mm
Earth radius, equatorial [a]	R_{eq}	6378.1367(1) km
Earth's radius, polar [a]	R_p	6356.7517(1) km
Equator–pole distance [a]		10 001.966 km (average)
Earth's flattening [a]	e	1/298.25231(1)
Earth's av. density	ρ	5.5 Mg/m^3
Earth's age		4.55 Ga = 143 Ps
Moon's radius	R_{mv}	1738 km in direction of Earth
Moon's radius	R_{mh}	17.. km in other two directions
Moon's mass	M_m	$7.35 \cdot 10^{22}$ kg
Moon's mean distance [b]	d_m	384 401 km
Moon's perigeon		typically 363 Mm, hist. minimum 359 861 km
Moon's apogeon		typically 404 Mm, hist. maximum 406 720 km
Moon's angular size [c]		avg. 0.5181° = 31.08′, min. 0.49°, max. 0.55°
Moon's av. density	ρ	3.3 Mg/m^3
Sun's mass	$M_\odot$	$1.988\,43(3) \cdot 10^{30}$ kg
Sun's grav. length	$l_\odot = 2GM_\odot/c^2$	2.953 250 08 km
Sun's luminosity	$L_\odot$	384.6 YW
solar radius, equatorial	$R_\odot$	695.98(7) Mm

Quantity	Name	Value
Sun's angular size		0.53° average; minimum on fourth of July (aphelion) 1888″, maximum on fourth of January (perihelion) 1952″
suns's av. density	ρ	$1.4\ \mathrm{Mg/m^3}$
Sun's distance, average	AU	149 597 870.691(30) km
Sun's age		4.6 Ga
solar velocity around centre of galaxy	$v_{\odot g}$	220(20) km/s
solar velocity against cosmic background	$v_{\odot b}$	370.6(5) km/s
distance to galaxy centre		8.0(5) kpc = 26.1(1.6) kal
Milky Ways's age	Abell	13.6 Ga
Milky Ways's size		*c.* 10^{21} m or 100 kal
Milky Ways's mass		10^{12} solar masses, *c.* $2 \cdot 10^{42}$ kg
Jupiter's mass	M	$1.90 \cdot 10^{27}$ kg
Jupiter's radius, equatorial	R	71.398 Mm
Jupiter's radius, polar	R	67.1(1) Mm
Jupiter's average distance from Sun	D	778 412 020 km
most distant galaxy	1835 IR1916	$13.2 \cdot 10^9$ al $= 1.25 \cdot 10^{26}$ m, red-shift 10

a. The shape of the Earth is described most precisely with the World Geodetic System. The last edition dates from 1984. For an extensive presentation of its background and its details, see the http://www.eurocontrol.be/projects/eatchip/wgs84/start.html website. The International Geodesic Union has refined the data in 2000. The radii and the flattening given here are those for the 'mean tide system'. They differ from those of the 'zero tide system' and other systems by about 0.7 m. The details are a science by its own.
b. Measured centre to centre. To know the precise position of the Moon at a given date, see the http://www.fourmilab.ch/earthview/moon-ap-per.html page, whereas for the planets see the page http://www.fourmilab.ch/solar/solar.html and the other pages on the site.
c. Angles are defined as follows: 1 degree = $1° = \pi/180$ rad, 1 (first) minute = $1' = 1°/60$, 1 second (minute) = $1'' = 1'/60$. The ancient units 'third minute' and 'fourth minute', each 1/60th of the preceding, are not in use any more. ('Minute' originally means 'very small', as it still does in modern English.)

Useful numbers

π	$3.14159\,26535\,89793\,23846\,26433\,83279\,50288\,41971\,69399\,37510_5$
e	$2.71828\,18284\,59045\,23536\,02874\,71352\,66249\,77572\,47093\,69995_9$
γ	$0.57721\,56649\,01532\,86060\,65120\,90082\,40243\,10421\,59335\,93992_3$
$\ln 2$	$0.69314\,71805\,59945\,30941\,72321\,21458\,17656\,80755\,00134\,36025_5$
$\ln 10$	$2.30258\,50929\,94045\,68401\,79914\,54684\,36420\,76011\,01488\,62877_2$
$\sqrt{10}$	$3.16227\,76601\,68379\,33199\,88935\,44432\,71853\,37195\,55139\,32521_6$

Ref. 1153

If the number π were *normal*, i.e. if all digits and digit combinations would appear with the same probability, then every text written or to be written, as well as every word spoken or to be spoken, can be found coded in its sequence. The property of normality has not yet been proven, even though it is suspected to be true. Does this mean that all wisdom is encoded in the simple circle? No. The property is nothing special, as it also applies to the number 0.123456789101112131415161718192021... and many others. Can you specify a few others?

Challenge 1484 n

By the way, in the graph of the exponential function e^x, the point $(0, 1)$ is the only point with two rational coordinates. If you imagine to paint in blue all points on the plane with two rational coordinates, the plane would look quite bluish. Nevertheless, the graph goes only through one of these points and manages to avoid all the others.

Bibliography

1132 Le Système International d'Unités, Bureau International des Poids et Mesures, Pavillon de Breteuil, Parc de Saint Cloud, 92 310 Sèvres, France. All new developments concerning SI units are published in the journal *Metrologia*, edited by the same body. Showing the slow pace of an old institution, the BIPM was on the internet only in 1998; it is now reachable on its simple site at http://www.bipm.fr. See also the http://www.utc.fr/~tthomass/Themes/Unites/index.html web site; it includes the biographies of men who gave name to various units. The site of its British equivalent, http://www.npl.co.uk/npl/reference/si_units.html, is much better; it provides many details as well as the English language version of the SI unit definitions. Cited on page 1060.

1133 The bible in the field of time measurement are the two volumes by J. Vanier & C. Audoin, *The Quantum Physics of Atomic Frequency Standards*, Adam Hilge, 1989. A popular account is Tony Jones, *Splitting the Second*, Institute of Physics Publishing, 2000.

The site http://opdaf1.obspm.fr/www/lexique.html gives a glossary of terms used in the field. On length measurements, see ... On mass and atomic mass measurements, see page 286. On electric current measurements, see ... On precision temperature measurements, see page 246. Cited on page 1060.

1134 The unofficial prefixes have been originally proposed in the 1990s by Jeff K. Aronson, professor at the University of Oxford, and might come into general usage in the future. Cited on page 1062.

1135 G.J. Stoney, On the Physical Units of Nature, *Philosophical Magazine* **11**, pp. 381–391, 1881. Cited on page 1065.

1136 D.J. Bird & al., Evidence for correlated changes in the spectrum and composition of cosmic rays at extremely high energies, *Physical Review Letters* **71**, pp. 3401–3404, 1993. Cited on page 1065.

1137 P.J. Hakonen & al., Nuclear antiferromagnetism in Rhodium metal at positive and negative nanokelvin temperature, *Physical Review Letters* **70**, pp. 2818–2821, 1993. See also his article in the *Scientific American*, January 1994. Cited on page 1066.

1138 G. Charpak & R.L. Garwin, The DARI, *Europhysics News* **33**, pp. 14–17, January/February 2002. Cited on page 1066.

1139 Measured values and ranges for physical quantities are collected in Horst Völz & Peter Ackermann, *Die Welt in Zahlen*, Spektrum Akademischer Verlag, 1996. Cited on page 1066.

1140 See e.g. K. Codling & L.J. Frasinski, Coulomb explosion of simple molecules in intense laser fields, *Contemporary Physics* **35**, pp. 243–255, 1994. Cited on page 1067.

1141 A. Zeilinger, The Planck stroll, *American Journal of Physics* **58**, p. 103, 1990. Can you find another example? Cited on page 1067.

Challenge 1485 e

1142 The most precise clock ever built is ... Cited on page 1067.

1143 J. Bergquist, editor, *Proceedings of the Fifth Symposium on Frequency Standards and Metrology*, World Scientific, 1997. Cited on page 1067.

1144 About short lifetime measurements, see e.g. the paper on the D meson lifetime, ... Cited on page 1067.

1145 About the long life of tantalum 180, see D. Belic & al., Photoactivation of $^{180}\mathrm{Ta}^{\mathrm{m}}$ and its implications for the nucleosynthesis of nature's rarest naturally occurring isotope, *Physical Review Letters* **83**, pp. 5242–5245, 20 December 1999. Cited on page 1068.

1146 About the detection of gravitational waves, see ... Cited on page 1068.

1147 See the clear and extensive paper by G.E. Stedman, Ring laser tests of fundamental physics and geophysics, *Reports on Progress of Physics* **60**, pp. 615–688, 1997. Cited on page 1068.

1148 Following a private communication by Richard Rusby, this is the value of 1997, whereas it was estimated as 99.975°C in 1989, as reported by Gareth Jones & Richard Rusby, Official: water boils at 99.975°C, *Physics World*, pp. 23–24, September 1989, and R.L. Rusby, Ironing out the standard scale, *Nature* **338**, p. 1169, March 1989. For more details on temperature measurements, see page 246. Cited on page 1068.

1149 See Newton's apples fall from grace, *New Scientist*, p. 5, 6 September 1996. More details can be found in R.G. Keesing, The history of Newton's apple tree, *Contemporary Physics* **39**, pp. 377–391, 1998. Cited on page 1068.

1150 The various concepts are even the topic of a separate international standard, ISO 5725, with the title *Accuracy and precision of measurement methods and results*. A good introduction is the book with the locomotive hanging out the window as title picture, namely John R. Taylor, *An Introduction to Error Analysis: the Study of Uncertainties in Physical Measurements*, 2nd edition, University Science Books, Sausalito, 1997. Cited on page 1069.

1151 P.J. Mohr & B.N. Taylor, *Reviews of Modern Physics* **59**, p. 351, 2000. This is the set of constants resulting from an international adjustment and recommended for international use by the Committee on Data for Science and Technology (CODATA), a body in the International Council of Scientific Unions, which regroups the International Union of Pure and Applied Physics (IUPAP), the International Union of Pure and Applied Chemistry (IUPAC) and many more. The IUPAC has a horrible website at http://chemistry.rsc.org/rsc/iupac.htm. Cited on page 1069.

1152 The details are given in the well-known astronomical reference, P. Kenneth Seidelmann, *Explanatory Supplement to the Astronomical Almanac*, 1992. Cited on page 1073.

1153 For information about the number π, as well as about other constants, the web address http://www.cecm.sfu.ca/pi/pi.html provides lots of data and references. It also has a link to the pretty overview paper on http://www.astro.virginia.edu/~eww6n/math/Pi.html and to many other sites on the topic. Simple formulae for π are

$$\pi + 3 = \sum_{n=1}^{\infty} \frac{n\,2^n}{\binom{2n}{n}} \tag{791}$$

or the beautiful formula discovered in 1996 by Bailey, Borwein and Plouffe

$$\pi = \sum_{n=0}^{\infty} \frac{1}{16^n} \Big(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \Big) . \qquad (792)$$

The site also explains the newly discovered methods to calculate specific binary digits of π without having to calculate all the preceding ones. By the way, the number of (consecutive) digits known in 1999 was over 1.2 million million, as told in *Science News* **162**, 14 December 2002. They pass all tests for a randomness, as the http://www.ast.univie.ac.at/~wasi/PI/pi_normal.html website explains. However, this property, called *normality*, has never been proven; it is the biggest open question about π.It is possible that the theory of chaotic dynamics will lead to a solution of this puzzle in the coming years.

Another method to calculate π and other constants was discovered and published by D.V. CHUDNOVSKY & G.V. CHUDNOVSKY, The computation of classical constants, Proceedings of the National Academy of Sciences (USA), **86**, pp. 8178–8182, 1989. The Chudnowsky brothers have built a supercomputer in Gregory's apartment for about 70 000 Euro, and for many years held the record for the largest number of digits for π. They battle already for decades with Kanada Yasumasa, who holds the record in 2000, calculated on an industrial supercomputer. New formulae to calculate π are still irregularly discovered.

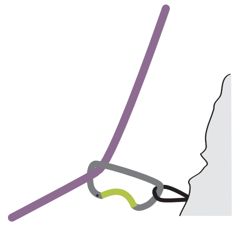
For the calculation of Euler's constant γ see also D.W. DETEMPLE, A quicker convergence to Euler's constant, *The Mathematical Intelligencer*, pp. 468–470, May 1993.

Challenge 1486 r

Note that little is known about properties of numbers; for example, it is still not known whether $\pi + e$ is a rational number or not! (It is believed that it is not.) Do you want to become a mathematician? Cited on page 1074.

Challenge 1487 n

Appendix C



Particle Properties

The following table lists the known and predicted elementary particles. There have been no changes in the list since the mid 1970s.

Table 84 The elementary particles

Radiation	electromagnetic interaction	weak interaction	strong interaction
	$\boxed{\gamma}$	$\boxed{W^+}$, $\boxed{W^-}$ $\boxed{Z^0}$	$\boxed{g_1}$ … $\boxed{g_8}$
	photon	intermediate vector bosons	gluons
	Radiation particles are bosons with spin 1. W^- is antiparticle of W^+; all others are their own antiparticles.		
Matter	generation 1	generation 2	generation 3
Leptons	$\boxed{e}$ $\boxed{\nu_e}$	$\boxed{\mu}$ $\boxed{\nu_\mu}$	$\boxed{\tau}$ $\boxed{\nu_\tau}$
Quarks (each in three colours)	$\boxed{d}$ $\boxed{u}$	$\boxed{s}$ $\boxed{c}$	$\boxed{t}$ $\boxed{b}$
	Matter particles are fermions with spin 1/2; all have a corresponding antiparticle.		
Hypothetical matter and radiation			
Higgs boson	$\boxed{H}$	predicted to have spin 0 and to be elementary	
Supersymmetric partners	$\boxed{se}$ … $\boxed{h}$	one partner for each of the above	

The following table contains the *complete list of properties* of all elementary particles. The future should not change it much.* The header of this table therefore lists the *complete* set of properties, after the quantum number of colour is added, which characterise any particle. This list thus also allows to deduce a complete characterization of the intrinsic properties of any *composed* moving entity, be it an object or an image.

Table 85 Elementary particle properties

Particle	Mass m	Lifetime τ or energy width, main decay modes	Isospin I, spin J, parity P, charge parity C	Charge Q, isospin I, strangeness S, charm C, beauty B, topness T	Lepton number L, baryon number B, R-parity
elementary radiation (bosons)					
photon γ	$0 \ (< 6 \cdot 10^{-16}\,\mathrm{eV}/c^2)$	stable	$I(J^{PC}) = 0, 1(1^{--})$	000 000	0, 0, 1
$W^{\pm}$	$80.75(64)\ \mathrm{GeV}/c^2$	$2.06(6)$ GeV 67.8(1.0)% hadrons, 32.1(2,0)% $l^+\nu$	$J = 1$	±100 000	0, 0, 1
Z	$91.187(7)\ \mathrm{GeV}/c^2$	$2.65(1) \cdot 10^{-25}$ s 69.90(15)% hadrons	$J = 1$	000 000	0, 0, 1
gluon	0	stable	$I(J^P) = 0(1^-)$	000 000	0, 0, 1
elementary matter (fermions): leptons					
electron e	$9.109\,381\,88(72) \cdot 10^{-31}\,\mathrm{kg} = 81.871\,0414(64)\ \mathrm{pJ}/c^2$ $= 0.510\,998\,902(21)\ \mathrm{MeV}/c^2 = 0.000\,548\,579\,9110(12)$ u gyromagnetic ratio $\mu_e/\mu_B = -1.001\,159\,652\,1883(42)$ electric dipole moment $d = (-0.3 \pm 0.8) \cdot 10^{-29}\, e$ m	$> 13 \cdot 10^{30}$ s	$J = \frac{1}{2}$	−100 000	1, 0, 1
muon μ	$0.188\,353\,109(16)$ yg $= 105.658\,3568(52)\ \mathrm{MeV}/c^2 = 0.113\,428\,9168(34)$ u gyromagnetic ratio $\mu_\mu/(e\hbar/2m_\mu) = -1.001\,165\,916\,02(64)$	$2.197\,03(4)$ µs 99 % $e^-\bar{\nu}_e\nu_\mu$	$J = \frac{1}{2}$	−100 000	1, 0, 1

* The official reference for all this data, worth a look by every physicist, is the massive collection by the particle data group, with the website http://pdg.web.cern.ch/pdg containing the most recent information. A printed review is published about every two years with updated data in one of the large journals on elementary particle physics. See for example S. Eidelman & al., Review of Particle Physics, *Physics Letters B* **592**, p. 1, 2004. For measured properties of these particles, the official reference is the set of CODATA values. The most recent list was published by P.J. Mohr & B.N. Taylor, *Reviews of Modern Physics* **59**, p. 351, 2000.

Particle	Mass m	Lifetime τ or energy width, main decay modes	Isospin I, spin J, parity P, charge parity C	Charge Q, isospin I, strangeness S, charm C, beauty B, topness T	Lepton number L, baryon number B, R-parity
	electric dipole moment $d = (3.7 \pm 3.4) \cdot 10^{-22}\,e\,\mathrm{m}$				
tau τ	$1.777\,05(29)\ \mathrm{GeV}/c^2$	$290.0(1.2)$ fs	$J = \frac{1}{2}$	$-100\,000$	$1, 0, 1$
el. neutrino ν_e	$< 7.2\ \mathrm{eV}/c^2$		$J = \frac{1}{2}$		$1, 0, 1$
muon neutrino ν_μ	$< 0.17\ \mathrm{MeV}/c^2$		$J = \frac{1}{2}$		$1, 0, 1$
tau neutrino ν_τ	$< 24\ \mathrm{MeV}/c^2$		$J = \frac{1}{2}$		$1, 0, 1$
elementary matter (fermions): quarks					
up quark u	$1.5 - 5\ \mathrm{MeV}/c^2$	see proton	$I(J^P) = \frac{1}{2}(\frac{1}{2}^+)$	$\frac{2}{3}\frac{1}{2}0000$	$0, 1/3, 1$
down quark d	$3 - 9\ \mathrm{MeV}/c^2$	see proton	$I(J^P) = \frac{1}{2}(\frac{1}{2}^+)$	$-\frac{1}{3}-\frac{1}{2}0000$	$0, 1/3, 1$
strange quark s	$60 - 170\ \mathrm{MeV}/c^2$		$I(J^P) = 0(\frac{1}{2}^+)$	$-\frac{1}{3}0-1000$	$0, 1/3, 1$
charm quark c	$1.25(15)\ \mathrm{GeV}/c^2$		$I(J^P) = 0(\frac{1}{2}^+)$	$\frac{2}{3}00+100$	$0, 1/3, 1$
bottom quark b	$4.25(15)\ \mathrm{GeV}/c^2$	$\tau = 1.33(11)$ ps	$I(J^P) = 0(\frac{1}{2}^+)$	$-\frac{1}{3}000{-}10$	$0, 1/3, 1$
top quark t	$173.8(5.2)\ \mathrm{GeV}/c^2$		$I(J^P) = 0(\frac{1}{2}^+)$	$\frac{2}{3}0000+1$	$0, 1/3, 1$
hypothetical, maybe elementary (boson)					
Higgs* H	$135(20)\ \mathrm{GeV}/c^2$		$J = 0$		
hypothetical elementary radiation (bosons)					
Selectron*			$J = 0$		$R = -1$
Smuon*			$J = 0$		$R = -1$
Stauon*			$J = 0$		$R = -1$
Sneutrinos*			$J = 0$		$R = -1$
Squark*			$J = 0$		$R = -1$
hypothetical elementary matter (fermions)					
Higgsino(s)*			$J = \frac{1}{2}$		$R = -1$
Wino* (a chargino)			$J = \frac{1}{2}$		$R = -1$
Zino* (a neutralino)			$J = \frac{1}{2}$		$R = -1$

Particle	Mass m	Lifetime τ or energy width, main decay modes	Isospin I, spin J, parity P, charge parity C	Charge Q, isospin I, strangeness S, charm C, beauty B, topness T	Lepton number L, baryon number B, R-parity
Photino*			$J = \frac{1}{2}$		$R = -1$
Gluino*			$J = \frac{1}{2}$		$R = -1$

Notes:

* Presently a hypothetical particle.

- To keep the table short, the header does not explicitly mention *colour*, the charge of the strong interactions. It has to be added to the list of object properties.
- 'Beauty' is now commonly called bottomness; similarly, 'truth' is now commonly called *topness*.
- Ref. 1154 The electron radius is below 10^{-22} m.
- Ref. 1155 It is possible to store single electrons in traps for many months.
- See also the table of SI prefixes on page 1061. About the eV/c^2 mass unit, see page 1065.
- Quantum numbers containing the word 'parity' are multiplicative; all others are additive.
- Time parity T, better called motion inversion parity, is equal to CP.
- The isospin I or I_Z is defined only for up and down quarks and their composites, such as the proton and the neutron. In the literature one also sees references to the so-called G-parity, defined as $G = (-1)^{IC}$.
- R-parity is a quantum number important in supersymmetric theories; it is related to the lepton number L, the baryon number B and the spin J through the definition $R = (-1)^{3B+L+2J}$. All particles from the standard model are R-even, whereas their superpartners are odd.
- The sign of the quantum numbers I_Z, S, C, B, T can be defined in different ways. In the standard assignment shown here, the sign of each of the non-vanishing quantum numbers is given by the sign of the charge of the corresponding quark.
- There is a difference between the half-life $t_{1/2}$ and the lifetime τ of a particle; the *half-life* is given by $t_{1/2} = \tau \ln 2$, where $\ln 2 \approx 0.693\,147\,18$; it is thus shorter than the lifetime. The *energy width* Γ of a particle is related to its lifetime τ by the indeterminacy relation $\Gamma\tau = \hbar$.
- The unified *atomic mass unit* is defined as (1/12) of the mass of an Carbon atom of the isotope ^{12}C at rest and in its ground state. One has $1\,\mathrm{u} = \frac{1}{12} m(^{12}\mathrm{C}) = 1.660\,5402(10)\,\mathrm{yg}$.
- See page 866 for the precise definition and meaning of the quark masses.

Using the table of elementary particle properties, together with the standard model and the fundamental constants, in principle *all* properties of composite matter and radiation can be deduced, including all those encountered in everyday life. (Can you explain how Challenge 1488 n the size of an object follows from them?) In a sense, this table contains the complete results of the study of matter and radiation, such as materials science, chemistry and biology.

The most important examples of composites are grouped in the following table.

Table 86 Properties of selected composites

Composite	Mass m, quantum numbers	Lifetime τ, main decay modes	Size (diam.)
mesons (hadrons, bosons) out of the over 130 types known			
pion $\pi^0(u\bar{u} - d\bar{d})/\sqrt{2}$	134.976 4(6) MeV/c^2 $I^G(J^{PC}) = 1^-(0^{-+})$, $S = C = B = 0$	84(6) as, 2γ 98.798(32)%	~ 1 fm
pion $\pi^+(u\bar{d})$	139.569 95(35) MeV/c^2 $I^G(J^P) = 1^-(0^-)$, $S = C = B = 0$	26.030(5) ns, $\mu^+\nu_\mu$ 99.987 7(4)%	~ 1 fm
kaon K^o_S	$m_{K^o_S}$	89.27(9) ps	
kaon K^o_L	$m_{K^o_S}$ + 3.491(9) μeV/c^2	51.7(4) ns	
kaon $K^\pm$ ($u\bar{s}$, $\bar{u}s$)	493.677(16) MeV/c^2	12.386(24) ns, $\mu^+\nu_\mu$ 63.51(18)% $\pi^+\pi^o$ 21.16(14)%	
kaon K^o ($d\bar{s}$) (50 %K_S, 50 % K_L)	497.672(31) MeV/c^2	n.a.	~ 1 fm
kaons $K^\pm$, K^o, K^o_S, K^o_L	$I(J^P) = \frac{1}{2}(0^-)$, $S = \pm 1$, $B = C = 0$		
baryons (hadrons, fermions) out of the over 100 types known			
proton p or N^+ (uud)	1.672 621 58(13) yg = 1.007 276 466 88(13) u = 938.271 998(38) MeV/c^2 $I(J^P) = \frac{1}{2}(\frac{1}{2}^+)$, $S = 0$ gyromagnetic ratio μ_p/μ_N = 2.792 847 337(29) electric dipole moment $d = (-4 \pm 6) \cdot 10^{-26}\,e$ m electric polarizability $\alpha = 12.1(0.9) \cdot 10^{-4}$ fm^3 magnetic polarizability $\alpha = 2.1(0.9) \cdot 10^{-4}$ fm^3	$\tau_{total} > 1.6 \cdot 10^{25}$ a, $\tau(p \rightarrow e^+\pi^o)$ >5.5 · 10^{32} a	0.89(1) fm Ref. 1156
neutron nor N^o (udd)	1.674 927 16(13) yg = 1.008 664 915 78(55) u = 939.565 330(38) MeV/c^2 $I(J^P) = \frac{1}{2}(\frac{1}{2}^+)$, $S = 0$ gyromagnetic ratio μ_n/μ_N = −1.913 042 72(45) electric dipole moment $d_n = (-3.3 \pm 4.3) \cdot 10^{-28}\,e$ m electric polarizability $\alpha = 0.98(23) \cdot 10^{-3}$ fm^3	887.0(2.0) s, $pe^-\bar{\nu}_e$ 100 %	~ 1 fm
omega Ω^- (sss)	1672.43(32) MeV/c^2 gyromagnetic ratio μ_Ω/μ_N = −1.94(22)	82.2(1.2) ps, ΛK^- 67.8(7)%, $\Xi^o\pi^-$ 23.6(7)%	~ 1 fm
composite radiation: glueballs			
glueball $f_0(1500)$	1503(11) MeV $I^G(J^{PC}) = 0^+(0^{++})$	full width 120(19) MeV	~ 1 fm
atoms out of the 114 known elements with over 2000 known nuclides		Ref. 1157	

Composite	Mass m, quantum numbers	Lifetime τ, main decay modes	Size (diam.)
hydrogen (^{1}H) [lightest]	1.007 825 032(1) u = 1.6735 yg		$2 \cdot 30$ pm
antihydrogen	1.007 u = 1.67 yg		$2 \cdot 30$ pm
helium (^{4}He) [smallest]	4.002 603250(1) u = 6.6465 yg		$2 \cdot 32$ pm
carbon (^{12}C)	12 u = 19.926 482(12) yg		$2 \cdot 77$ pm
bismuth ^{209}Bi [shortest living and rarest]	209 u	0.1 ps Ref. 1159	
tantalum ^{180}Ta [second longest living radioactive]	180 u	$> 10^{15}$ a Ref. 1158	
bismuth ^{209}Bi [longest living radioactive]	209 u	$1.9(2)10^{19}$ a Ref. 1159	
francium [largest of all]	223 u	...	$2 \cdot 0.28$ nm
atom 116 [heaviest of all]	289 u	...	
molecules out of the over 10^7 known types			
hydrogen (H_2)	~ 2 u	$> 10^{25}$ a	
water (H_2O)	~ 18 u	$> 10^{25}$ a	
ATP (adenosinetriphosphate)	507 u	$> 10^{10}$ a	*c.* 3 nm
human Y chromosome	$70 \cdot 10^6$ base pairs	$> 10^6$ a	*c.* 50 mm (uncoiled)
other composites			
blue whale nerve cell	~ 1 kg	~ 50 a	20 m
cell (red blood)	0.1 ng	7/120 days	~ 10 µm
cell (sperm)	10 pg	not fecundated: ~ 5 d	length 60 µm, head 3 µm times 5 µm
cell (ovule)	1 µg	fecundated: over 4000 million years	~ 120 µm
cell (E. Coli)	1 pg	4000 million years	body: 2 µm
adult human	$35\,\text{kg} < m < 350\,\text{kg}$	$\tau \approx 2.5 \cdot 10^9$ s Ref. 1160 ≈ 600 million breaths ≈ 2 500 million heartbeats < 122 a, 60 % H_2O and 40 % dust	~ 1.7 m
Heaviest living thing: aspen trees	$6.6 \cdot 10^6$ kg	> 130 a	> 4 km
larger composites	see the table on page 169.		

Notes (see also those of the previous table)

- G parity is defined only for mesons and given by $G = (-1)^{L+S+I} = (-1)^I \cdot C$.

- In 2003, experiments provided candidates for tetraquarks, namely the X(3872), Ds(2317), f(980) and Ds(2460), and for pentaquarks, namely the $\Theta_{+}(1500)$ particle. The future will show whether these interpretations are correct. Ref. 1161
- Neutrons bound in nuclei have a lifetime of at least 10^{20} years.
- The $f_0(1500)$ resonance is a candidate for the glueball ground state and thus for a *radiation composite*.
- The $Y(3940)$ resonance is a candidate for a hybrid meson, a composite of a gluon and a quark-antiquark pair. This prediction of 1980 seems to have been confirmed in 2005.
- In 2002, first evidence for the existence of tetra-neutrons was published by a French group. However, more recent investigations seem to have refuted the claim. Ref. 1162
- The number of existing molecules is several orders of magnitude larger than the number of analysed and listed molecules.
- Some nuclei are not yet discovered; in 2002 the known nuclei range from 1 to 116, but 113, 115 and 117 are still missing.
- The first *anti-atoms*, made of antielectrons and antiprotons, have been made in January 1996 at CERN in Geneva. All properties for antimatter checked so far are consistent with the predictions. Ref. 1163
- The charge parity C is defined only for certain neutral particles, namely those which are different from their antiparticles. For neutral mesons the charge parity is given by $C = (-1)^{L+S}$, where L is the orbital angular momentum.
- P is the parity under space inversion $\mathbf{r} \rightarrow -\mathbf{r}$. For mesons, it is connected through $P = (-1)^{L+1}$ with the orbital angular momentum L.
- The electric polarizability is defined on page 548; it is predicted to vanish for all elementary particles.

The most important matter composites are the *atoms*. Their size, structure and interactions determine the properties and colour of everyday objects. Atom types, also called *elements* in chemistry, are most efficiently grouped in the so-called *periodic table*, in which atoms behaving similarly are neighbours. The table results from the various ways that protons, neutrons and electrons can combine to form aggregates.

There are similar tables also for the mesons (made of any two quarks) and the baryons (made of three quarks). Neither the meson nor the baryon table is included here; they can be found in the cited *Review of Particle Physics*. In fact, the baryon table still has a number of vacant spots. However, the missing particles are extremely heavy and short lived (which means expensive to make and detect) and their discovery is not expected to yield deep new insights.

The atomic number gives the number of protons and of electrons found in an atom of a given element. This number also specifies each element in its chemical behaviour. Most – but not all – elements up to 92 are found on Earth; the others can be produced in laboratories. The last element discovered is element 116. (In a famous case of research fraud, a scientist in the 1990s tricked two whole research groups into claiming to have made and observed elements 116 and 118. Element 116 was independently made and observed by another group later on.) Nowadays, extensive physical and chemical data is available for every element. Ref. 1166

Elements in the same *group* behave similarly in chemical reactions. The *periods* define the repetition of these similarities. More elaborate periodic tables can be found on the http://chemlab.pc.maricopa.edu/periodic website. The most beautiful of them all can be found on page 781 of this text.

Table 87 The periodic table of the elements known in 2005, with their atomic numbers

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	I	II	IIIa	IVa	Va	VIa	VIIa		VIIIa		Ia	IIa	III	IV	V	VI	VII	VIII
Period																		
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Uuu	112 Uub	113 Uut	114 Uuq	115 Uup	116 Uuh	117	118

Lanthanoids	*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
Actinoids	**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Group 1 are the alkali metals (though hydrogen is a gas), group 2 the Earth-alkali metals. Actinoids, lanthanoids and groups 3 to 13 are metals; in particular, groups 3 to 12 are transition or heavy metals. The elements of group 16 are called *chalkogens*, i.e. ore-formers; group 17 are the *halogens*, i.e. the salt-formers, and group 18 are the inert *noble gases*, which do not form (almost) any chemical compounds. The groups 13, 14 and 15 contain metals, semimetals, a liquid and gases; they have no special name. Groups 1 and 13 to 17 are central for the chemistry of (animal) life; in fact, 96 % of living matter is made of C, O, N, H;* almost 4 % of P, S, Ca, K, Na, Cl; trace elements such a Mg, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Cd, Pb, Sn, Li, Mo, Se, Si, I, F, As, B form the rest. That makes over 30

* The average formula of life is approximately $C_5H_{40}O_{18}N$.

elements known to be essential for animal life. The full list is not yet known; candidate elements to extend this list are Al, Br, Ge and W.

Many elements exist in versions with different numbers of neutrons in their nucleus and thus with different mass; these various *isotopes* – called this way because they are found on the *same place* in the periodic table – behave identically in chemical reactions. There are over 2000 of them.

Ref. 1157, Ref. 1164

Table 88 The elements and their main properties

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Actinium[b]	Ac	89	(227.0277(1)) 21.77(2) a	(188)	highly radioactive metallic rare Earth (Greek 'aktis' ray) 1899, used as alpha-emitting source
Aluminium	Al	13	26.981 538(2) stable	118c, 143m	light metal (Latin 'alumen' for alum) 1827, used in machine construction and living beings
Americium[b]	Am	95	(243.0614(1)) 7.37(2) ka	(184)	radioactive metal (Italian 'America') 1945, used in smoke detectors
Antimony	Sb	51	121.760(1)[f] stable	137c, 159m, 205v	toxic semimetal (via Arabic from Latin stibium, itself from Greek, Egyptian for one of its minerals) colours rubber, used in medicines, constituent of enzymes
Argon	Ar	18	39.948(1)[f] stable	(71n)	noble gas (Greek 'argos' inactive, from 'anergos' without energy) 1894, third component of air, used for welding and in lasers
Arsenic	As	33	74.921 60(2) stable	120c, 185v	poisonous semimetal (Greek 'arsenikon' tamer of males) antiquity, for poisoning pigeons and doping semiconductors
Astatine[b]	At	85	(209.9871(1)) 8.1(4) h	(140)	radioactive halogen (Greek 'astatos' unstable) 1940, no use
Barium	Ba	56	137.327(7) stable	224m	Earth-alkali metal (Greek 'bary' heavy) 1808, used in vacuum tubes, paint, oil industry, pyrotechnics and X-ray diagnosis
Berkelium[b]	Bk	97	(247.0703(1)) 1.4(3) ka	n.a.	made in lab, probably metallic (Berkeley, US town) 1949, due to its rarity no use
Beryllium	Be	4	9.012 182(3) stable	106c, 113m	toxic Earth-alkali metal (Greek 'beryllos', a mineral) 1797, used in light alloys, in nuclear industry as moderator
Bismuth	Bi	83	208.980 38(2) stable	170m, 215v	diamagnetic metal (Latin via German 'weisse Masse' white mass) 1753, used in magnets, alloys, fire safety, cosmetics, as catalyst, nuclear industry

Name	Sym-bol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Bohrium[b]	Bh	107	(264.12(1)) 0.44 s[g]	n.a.	made in lab, probably metallic (after Niels Bohr) 1981, found in reactions
Boron	B	5	10.811(7)[f] stable	83c	semimetal, semiconductor (Latin borax, from Arabic and Persian for brilliant) 1808, used in glass, bleach, pyrotechnics, rocket fuel, medicine
Bromine	Br	35	79.904(1) stable	120c, 185v	red-brown liquid (Greek 'bromos' strong odour) 1826, fumigants, photography, water purification, dyes, medicines
Cadmium	Cd	48	112.411(8)[f] stable	157m	heavy metal, cuttable, screaming (Greek kadmeia, a Zinc Carbonate mineral where it was discovered) 1817, electroplating, solder, batteries, TV phosphors, dyes
Caesium	Cs	55	132.905 45(2) stable	273m	alkali metal (Latin 'caesius' sky blue) 1860, getter in vacuum tubes, photoelectric cells, ion propulsion, atomic clocks
Calcium	Ca	20	40.078(4)[f] stable	197m	Earth-alkali metal (Latin 'calcis' chalk) antiquity, pure in 1880, found in stones and bones, reducing agent, alloying
Californium[b]	Cf	98	(251.0796(1)) 0.90(5) ka	n.a.	made in lab, probably metallic, strong neutron emitter (Latin 'calor' heat and 'fornicare' have sex, the land of hot sex :-) 1950, used as neutron source, for well logging
Carbon	C	6	12.0107(8)[f] stable	77c	makes up coal and diamond (Latin 'carbo' coal) antiquity, used to build most life forms
Cerium	Ce	58	140.116(1)[f] stable	183m	rare Earth metal (after asteroid Ceres, roman goddess) 1803, cigarette lighters, incandescent gas mantles, glass manufacturing, self cleaning ovens, carbon-arc lighting in the motion picture industry, catalyst, metallurgy
Chlorine	Cl	17	35.453(2)[f] stable	102c, 175v	green gas (Greek 'chloros' yellow-green) 1774, drinking water, polymers, paper, dyes, textiles, medicines, insecticides, solvents, paints, rubber.
Chromium	Cr	24	51.9961(6) stable	128m	transition metal (Greek 'chromos' colour) 1797, hardens steel, make steel stainless, alloys, electroplating, green glass dye, catalyst, etc.

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Cobalt	Co	27	58.933 200(9) stable	125m	ferromagnetic transition metal (German 'Kobold' goblin) 1694, part of vitamin B_{12}, magnetic alloys, high duty alloys, enamel dyes, ink, and animal nutrition.
Copper	Cu	29	63.546(3)[f] stable	128m	red metal (Latin cuprum from Cyprus island) antiquity, part of many enzymes, electric conductors, bronze, brass and other alloys, agaecides, etc.
Curium[b]	Cm	96	(247.0704(1)) 15.6(5) Ma	n.a.	highly radioactive, silver-coloured (after Pierre and Marie Curie) 1944, used as radioactivity source
Darmstadtium[b]	Ds	110	(271) 1.6 min[g]	n.a.	(after the German city) 1994, no use
Dubnium[b]	Db	105	(262.1141(1)) 34(5) s	n.a.	made in lab in small numbers, radioactive ('Dubna' Russian city) 1967, no use (once know as Hahnium)
Dysprosium	Dy	66	162.50(3)[f] stable	177m	rare Earth metal (Greek 'dysprositos' difficult to obtain) 1886, used in laser materials, as infrared source material, and in nuclear industry
Einsteinium[b]	Es	99	(252.0830(1)) 472(2) d	n.a.	made in lab, radioactive (after Albert Einstein) 1952, no use
Erbium	Er	68	167.259(3)[f] stable	176m	rare Earth metal ('Ytterby' Swedish town) 1843, used in metallurgy and optical fibres
Europium	Eu	63	151.964(1)[f] stable	204m	rare Earth metal (named after the continent) 1901, used in red screen phosphor for tv tubes
Fermium[b]	Fm	100	(257.0901(1)) 100.5(2) d	n.a.	(after Enrico Fermi) 1952, no use
Fluorine	F	9	18.998 4032(5) stable	62c, 147v	gaseous halogen (from fluorine, a mineral, from Greek 'fluo' flow) 1886, used in polymers and toothpaste
Francium[b]	Fr	87	(223.0197(1)) 22.0(1) min	(278)	radioactive metal (from France) 1939, no use
Gadolinium	Gd	64	157.25(3)[f] stable	180m	(after Johan Gadolin) 1880
Gallium	Ga	31	69.723(1) stable	125c, 141m	almost liquid metal (Latin for both the discoverer's name and his nation, France) 1875, used in optoelectronics
Germanium	Ge	32	72.64(1) stable	122c, 195v	semiconductor (from Germania, as opposed to gallium) 1886, used in electronics

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Gold	Au	79	196.966 55(2) stable	144m	heavy noble metal (Latin aurum) antiquity, electronics, jewels
Hafnium	Hf	72	178.49(2)[c] stable	158m	metal (Latin for Copenhagen) 1923, alloys, incandescent wire
Hassium[b]	Hs	108	(277) 16.5 min[g]	n.a.	radioactive element (Latin form of German state Hessen) 1984, no use
Helium	He	2	4.002 602(2)[f] stable	(31n)	noble gas (Greek 'helios' Sun) where it was discovered 1895, used in balloons, stars, divers gas and cryogenics
Holmium	Ho	67	164.930 32(2) stable	177m	metal (Stockholm, Swedish capital) 1878, alloys
Hydrogen	H	1	1.007 94(7)[f] stable	30c	reactive gas (Greek for water-former) 1766, used in building stars and universe
Indium	In	49	114.818(3) stable	141c, 166m	soft metal (Greek 'indikon' indigo) 1863, used in solders and photocells
Iodine	I	53	126.904 47(3) stable	140c, 198v	blue-black solid (Greek 'iodes' violet) 1811, used in photography
Iridium	Ir	77	192.217(3) stable	136m	precious metal (Greek 'iris' rainbow) 1804, contact layers
Iron	Fe	26	55.845(2) stable	127m	metal (Latin ferrum) antiquity, used in metallurgy
Krypton	Kr	36	83.80(1)[f] stable	(88n)	noble gas (Greek 'kryptos' hidden) 1898, used in lasers
Lanthanum	La	57	138.9055(2)[c,f] stable	188m	reactive rare Earth metal (Greek 'lanthanein' to be hidden) 1839, used in lamps and in special glasses
Lawrencium[b]	Lr	103	(262.110 97(1)) 3.6(3) h	n.a.	appears in reactions (after Ernest Lawrence) 1961, no use
Lead	Pb	82	207.2(1)[c,f] stable	175m	poisonous, malleable heavy metal (Latin plumbum) antiquity, used in car batteries, radioactivity shields, paints
Lithium	Li	3	6.941(2)[f] stable	156m	light alkali metal with high specific heat (Greek 'lithos' stone) 1817, used in batteries, anti-depressants, alloys and many chemicals
Lutetium	Lu	71	174.967(1)[f] stable	173m	rare Earth metal (Latin 'Lutetia' , old name of Paris) 1907, used as catalyst

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Magnesium	Mg	12	24.3050(6) stable	160m	light common alkaline Earth metal, (from Magnesia, a Greek district in Thessalia) 1755, used in alloys, pyrotechnics, chemical synthesis and medicine, found in chlorophyll
Manganese	Mn	25	54.938 049(9) stable	126m	brittle metal (Italian 'Manganese', a mineral) 1774, used in alloys, colours amethyst and permanganate
Meitnerium[b]	Mt	109	(268.1388(1)) 0.070 s[g]	n.a.	appears in nuclear reactions (after Lise Meitner) 1982, no use
Mendelevium[b]	Md	101	(258.0984(1)) 51.5(3) d	n.a.	appears in nuclear reactions (after Dimitri Ivanovitch Mendeleiev) 1955, no use
Mercury	Hg	80	200.59(2) stable	157m	liquid heavy metal (Greek 'hydrargyrum' liquid silver) antiquity, used in switches, batteries, lamps, amalgam alloys
Molybdenum	Mo	42	95.94(1)[f] stable	140m	metal (Greek 'molybdos' lead) 1788, used in alloys, as catalyst, in enzymes and lubricants
Neodymium	Nd	60	144.24(3)[c, f] stable	182m	(Greek 'neos' and 'didymos' new twin) 1885
Neon	Ne	10	20.1797(6)[f] stable	(36n)	noble gas (Greek 'neos' new) 1898, used in lamps, lasers and cryogenics
Neptunium[b]	Np	93	(237.0482(1)) 2.14(1) Ma	n.a.	radioactive metal (planet Neptune, after Uranus) 1940, appears in nuclear reactors, used in neutron detection and by the military
Nickel	Ni	28	58.6934(2) stable	125m	metal (German 'Nickel' goblin) 1751, used in coins, stainless steels, batteries, as catalyst
Niobium	Nb	41	92.906 38(2) stable	147m	ductile metal (Greek 'Niobe', mythical daughter of Tantalos) 1801, used in arc welding, alloys, jewelry, superconductors
Nitrogen	N	7	14.0067(2)[f] stable	70c, 155v	diatomic gas (Greek for nitre former) 1772, found in air, in living organisms, viagra, fertilizers, explosives
Nobelium[b]	No	102	(259.1010(1)) 58(5) min	n.a.	(after Alfred Nobel) 1958, no use
Osmium	Os	76	190.23(3)[f] stable	135m	heavy metal (from Greek 'osme' odour) 1804, used for fingerprint detection and in very hard alloys

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Oxygen	O	8	15.9994(3)[f] stable	66c, 152v	transparent, diatomic gas (formed from Greek to mean 'acid former') 1774, used for combustion, blood regeneration, to make most rocks and stones, in countless compounds, colours auroras red
Palladium	Pd	46	106.42(1)[f] stable	138m	heavy metal (from asteroid 'Pallas' after the Greek goddess) 1802, used in alloys, white gold, catalysts, for hydride storage
Phosphorus	P	15	30.973 761(2) stable	109c, 180v	poisonous, waxy, white solid (Greek 'phosphoros' light bearer) 1669, fertilizers, glasses, porcelain, steels and alloys, living organisms, bones
Platinum	Pt	78	195.078(2) stable	139m	silvery-white, ductile, noble heavy metal (Spanish 'platina' little silver) pre-columbian, again in 1735, used in corrosion-resistant alloys, magnets, furnaces, catalysts, fuel cells, cathodic protection systems for large ships and pipelines; being a catalyst, a fine platinum wire glows red hot when placed in vapour of methyl alcohol, an effect used hand warmers
Plutonium	Pu	94	(244.0642(1)) 80.0(9) Ma	n.a.	extremely toxic alpha emitting metal (after the planet) synthesized 1940, found in nature 1971, used as nuclear explosive, and to power space equipment, such as satellites and the measurement equipment brought to the Moon by the Apollo missions
Polonium	Po	84	(208.9824(1)) 102(5) a	(140)	alpha-emitting, volatile metal (from Poland) 1898, used as thermoelectric power source in space satellites, as neutron source when mixed with beryllium; used in the past to eliminate static charges in factories, and on brushes for removing dust from photographic films.
Potassium	K	19	39.0983(1) stable	238m	reactive, cuttable light metal (Latin 'kalium' from Arabic 'quilyi', a plant used to produce potash, German 'Pottasche') 1807, part of many salts and rocks, essential for life, used in fertilizers, essential to chemical industry

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Praeseodymium	Pr	59	140.907 65(2) stable	183m	white, malleable rare Earth metal (Greek 'praesos didymos' green twin) 1885, used in cigarette lighters, material for carbon arcs used by the motion picture industry for studio lighting and projection, glass and enamel dye, darkens welders goggles
Promethium[b]	Pm	61	(144.9127(1)) 17.7(4) a	181m	radioactive rare Earth (from the Greek mythical figure of Prometheus) 1945, used as beta source and to excite phosphors
Protactinium	Pa	91	(231.035 88(2)) 32.5(1) ka	n.a.	radioactive metal (Greek 'protos' first, as it decays into Actinium) 1917, found in nature, no use
Radium	Ra	88	(226.0254(1)) 1599(4) a	(223)	highly radioactive metal (Latin 'radius' ray) 1898, no use any more; once used in luminous paints and as radioactive source and in medicine
Radon	Rn	86	(222.0176(1)) 3.823(4) d	(130n)	radioactive noble gas (from its old name 'radium emanation') 1900, no use (any more), found in soil, produces lung cancer
Rhenium	Re	75	186.207(1)[c] stable	138m	(Latin 'rhenus' for Rhine river) 1925, used in filaments for mass spectrographs and ion gauges, superconductors, thermocouples, flash lamps, and as catalyst
Rhodium	Rh	45	102.905 50(2) stable	135m	white metal (Greek 'rhodon' rose) 1803, used to harden platinum and palladium alloys, for electroplating, and as catalyst
Roentgenium[b]	Rg	111	(272.1535(1)) 1.5 ms[g]	n.a.	1994, no use
Rubidium	Rb	37	85.4678(3)[f] stable	255m	silvery-white, reactive alkali metal (Latin 'rubidus' red) 1861, used in photocells, optical glasses, solid electrolytes
Ruthenium	Ru	44	101.107(2)[f] stable	134m	white metal (Latin 'Rhuthenia' for Russia) 1844, used in platinum and palladium alloys, superconductors, as catalyst; the tetroxide is toxic and explosive
Rutherfordium[b]	Rf	104	(261.1088(1)) 1.3 min[g]	n.a.	radioactive transactinide (after Ernest Rutherford) 1964, no use

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Samarium	Sm	62	150.36(3)[c,f] stable	180m	silver-white rare Earth metal (from the mineral Samarskite, after Wassily Samarski) 1879, used in magnets, optical glasses, as laser dopant, in phosphors, in high power light sources
Scandium	Sc	21	44.955 910(8) stable	164m	silver-white metal (from Latin 'Scansia' Sweden) 1879, the oxide is used in high-intensity mercury vapour lamps, a radioactive isotope is used as tracer
Seaborgium[b]	Sg	106	266.1219(1) 21 s[g]	n.a.	radioactive transurane (after Glenn Seaborg) 1974, no use
Selenium	Se	34	78.96(3)[f] stable	120c, 190v	red or black or grey semiconductor (Greek 'selene' Moon) 1818, used in xerography, glass production, photographic toners, as enamel dye
Silicon	Si	14	28.0855(3)[f] stable	105c, 210v	grey, shiny semiconductor (Latin 'silex' pebble) 1823, Earth's crust, electronics, sand, concrete, bricks, glass, polymers, solar cells, essential for life
Silver	Ag	47	107.8682(2)[f] stable	145m	white metal with highest thermal and electrical conductivity (Latin argentum, Greek 'argyros') antiquity, used in photography, alloys, to make rain,
Sodium	Na	11	22.989 770(2) stable	191m	light, reactive metal (Egyptian, Arabic 'natrium' and Arabic 'souwad' soda) component of many salts, soap, paper, soda, salpeter, borax, and essential for life
Strontium	Sr	38	87.62(1)[f] stable	215m	silvery, spontaneously igniting light metal (Strontian, Scottish town) 1790, used in TV tube glass, in magnets, and in optical materials
Sulphur	S	16	32.065(5)[f] stable	105c, 180v	yellow solid (Latin) antiquity, used in gunpowder, in sulphuric acid, rubber vulcanization, as fungicide in wine production, and is essential for life; some bacteria use sulphur instead of oxygen in their chemistry
Tantalum	Ta	73	180.9479(1) stable	147m	heavy metal (Greek Tantalos, a mythical figure) 1802, used for alloys, surgical instruments, capacitors, vacuum furnaces, glasses

Name	Sym-bol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Technetium[b]	Tc	43	(97.9072(1)) 6.6(10) Ma	136m	radioactive (Greek 'technetos 'artificial') 1939, used as radioactive tracer and in nuclear technology
Tellurium	Te	52	127.60(3)[f] stable	139c, 206v	brittle, garlic-smelling semiconductor (Latin 'tellus' Earth) 1783, used in alloys and as glass component
Terbium	Tb	65	158.925 34(2) stable	178m	malleable rare Earth metal ('Ytterby' Swedish town) 1843, used as dopant in optical material
Thallium	Tl	81	204.3833(2) stable	172m	soft, poisonous heavy metal (Greek 'thallos' branch) 1861, used as poison and for infrared detection
Thorium	Th	90	232.0381(1)[d,f] 14.0(1) Ga	180m	radioactive (nordic god Thor, as in thursday) 1828, found in nature, heats Earth, used as oxide, in alloys and as coating
Thulium	Tm	69	168.934 21(2) stable	175m	rare Earth metal (Thule, mythical name for Scandinavia) 1879, found in monazite
Tin	Sn	50	118.710(7)[f] stable	139c, 210v, 162m	grey metal that cries when bent (Latin stannum) antiquity, used in paint, bronze and superconductors
Titanium	Ti	22	47.867(1) stable	146m	metal (Greek hero Titanos) 1791, alloys, fake diamonds
Tungsten	W	74	183.84(1) stable	141m	heavy, highest melting metal (German Wolfram, Swedish 'tung sten' heavy stone) 1783, light bulbs
Ununbium[b]	Uub	112	(285) 15.4 min[g]	n.a.	1996, no use
Ununtrium	Uut	113		n.a.	2004, no use
Ununquadium[b]	Uuq	114	(289) 30.4 s[g]	n.a.	1999, no use
Ununpentium	Uup	115		n.a.	2004, no use
Ununhexium[b]	Uuh	116	(289) 0.6 ms[g]	n.a.	2000 (earlier claim was false), no use
Ununseptium	Uus	117		n.a.	not yet observed
Ununoctium	Uuo	118		n.a.	not yet observed, but false claim in 1999
Uranium	U	92	238.028 91(3)[d,f] $4.468(3) \cdot 10^9$ a	156m	radioactive and of high density (Planet Uranos, the Greek sky god) 1789, found in pechblende and other minerals, used for nuclear energy
Vanadium	V	23	50.9415(1) stable	135m	metal ('Vanadis' scandinavian goddess of beauty) 1830, used in steel
Xenon	Xe	54	131.293(6)[f] stable	(103n) 200v	noble gas (Greek 'xenos' foreign) 1898, used in lamps and lasers

Name	Symbol	At. n.	Av. mass[a] in u (error), longest lifetime	At.[e] radius in pm	Main properties (naming), discovery date and use
Ytterbium	Yb	70	173.04(3)[f] stable	174m	malleable heavy metal ('Ytterby' Swedish town) 1878, used in superconductors
Yttrium	Y	39	88.905 85(2) stable	180m	malleable light metal ('Ytterby' Swedish town) 1794, used in lasers
Zinc	Zn	30	65.409(4) stable	139m	heavy metal (German 'Zinke' protuberance) antiquity, iron rust protection
Zirconium	Zr	40	91.224(2)[f] stable	160m	heavy metal (from the mineral zircon, after Arabic 'zargum' golden colour) 1789, chemical and surgical instruments, nuclear industry

a. The atomic mass unit is defined as $1\,\mathrm{u} = \frac{1}{12} m(^{12}\mathrm{C}) = 1.660\,5402(10)\,\mathrm{yg}$. For elements found on Earth, the *average* atomic mass for the natural occurring isotope mixture is given, with the
Ref. 1164 error in the last digit in brackets. For elements not found on Earth, the mass of the *longest living* isotope is given; as it is not an average, it is written in brackets, as usual in this domain.
b. The element is not found on Earth due to its short lifetime.
c. The element contains at least one radioactive isotope.
d. The element has no stable isotopes.
e. The *atomic radius* does not exist. In fact, all atoms being clouds, they have no boundary. The size of atoms thus depends on the way it is defined. Several approximate definitions are possible. Usually, the radius is defined in a way to be useful for the estimation of distances between atoms.
Ref. 1165 This distance is different for different bond types. Radii for metallic bonds are labelled m, radii
Ref. 1165 for (single) covalent bonds with carbon c and Van der Waals radii v. Noble gas radii are denoted n. Note that values found in the literature vary by about 10 %; values in brackets lack literature references.

The covalent radius is often up to 0.1 nm smaller than the metallic radius for elements on the (lower) left of the periodic table; on the (whole) right side it is essentially equal to the metallic radius. In between, the difference between the two decreases towards the right. Are you able to
Challenge 1489 n explain why? By the way, ionic radii differ considerably from atomic ones and depend both on the ionic charge and the element itself.

All these values are for atoms in their ground state. Excited atoms can be hundreds of times larger than atoms in the ground state; however, excited atoms do not form solids or chemical compounds.
f. The isotopic composition and thus the average atomic mass of the element varies depending on the mining place or on subsequent human treatment, and can lie outside the values given. For example, the atomic mass of commercial lithium ranges between 6.939 and 6.996 u. The masses of isotopes are known in atomic mass units with nine or more significant digits, and usually with one
Ref. 1157 or two digits less in kilogram. The errors in the atomic mass are thus mainly due to the variations
Ref. 1164 in isotope composition.
g. The lifetime errors are asymmetric or not well known.

Bibliography

1154 The electron radius limit is deduced from the $g - 2$ measurements, as explained in the Nobel prize talk by HANS DEHMELT, Experiments with an isolated subatomic particle at rest, *Reviews of Modern Physics* **62**, pp. 525–530, 1990, or in HANS DEHMELT, Is the electron a composite particle?, *Hyperfine Interactions* **81**, pp. 1–3, 1993. Cited on page 1081.

1155 G. GABRIELSE, H. DEHMELT & W. KELLS, Observation of a relativistic, bistable hysteresis in the cyclotron motion of a single electron, *Physical Review Letters* **54**, pp. 537–540, 1985. Cited on page 1081.

1156 The proton charge radius was determined by measuring the frequency of light emitted by hydrogen atoms to high precision by TH. UDEM, A. HUBER, B. GROSS, J. REICHERT, M. PREVEDELLI, M. WEITZ & T.W. HAUSCH, Phase-coherent measurement of the hydrogen 1S-2S transition frequency with an optical frequency interval divider chain, *Physical Review Letters* **79**, pp. 2646–2649, 1997. Cited on page 1082.

1157 For a full list of isotopes, see R.B. FIRESTONE, *Table of Isotopes, Eighth Edition, 1999 Update*, with CDROM, John Wiley & Sons, 1999. For a list of isotopes on the web, the standard is the korean website by J. CHANG, http://hpngp01.kaeri.re.kr/CoN/index.html, mirrored at http://www.nes.ruhr-uni-bochum.de/CoN. For a list of precise isotope masses, see the http://csnwww.in2p3.fr website. Cited on pages 1082, 1086, and 1095.

1158 About the long life of tantalum 180, see D. BELIC & al., Photoactivation of $^{180}Ta^{m}$ and its implications for the nucleosynthesis of nature's rarest naturally occurring isotope, *Physical Review Letters* **83**, pp. 5242–5245, 20 December 1999. Cited on page 1083.

1159 The ground state of bismuth 209 was though to be stable until early 2003. It was then discovered that it was radioactive, though with a record lifetime, as reported by P. DE MARCILLAC & al., *Nature* **422**, p. 876, 2003. By coincidence, the excited state 83 MeV above the ground state of the same 209 bismuth nucleus is the shortest known radioactive nuclear state. Cited on page 1083.

1160 STEPHEN J. GOULD, 1980, The Panda's thumb, W.W. Norton & Co., New York. This is one of several interesting and informative books on evolutionary biology by the best writer in the field. Cited on page 1083.

1161 A tetraquark is seen as the best explanation for the $f_0(980)$ resonance at 980 MeV. However, this is still under investigation. Pentaquarks were first predicted by Maxim Polyakov, Dmitri Diakonov, and Victor Petrov in 1997. Two 2003 experiments claimed to confirm the existence with a mass of 1540 MeV; see K. HICKS, An experimental review of the Θ^+ pentaquark, http://www.arxiv.org/abs/hep-ex/0412048. Recent 2005 results however seem to rule out that the 1540 MeV particle is a pentaquark. Cited on page 1084.

1162 F. MARQUES & al., Detection of neutron clusters, *Physical Review C* **65**, p. 044006, 2002. Opposite results are by B.M. SHERRILL & C.A. BERTULANI, Proton-tetraneutron elastic scattering, *Physical Review C* **69**, p. 027601, 2004, and D.V. ALEKSANDROV & al., Search for resonances in the three- and four-neutron systems in the 7Li(7Li, 11C)3n and 7Li(7Li, 10C)4n reactions, *JETP Letters* **81**, p. 43, 2005. Cited on page 1084.

1163 For a good review, see the article by P.T. GREENLAND, Antimatter, *Contemporary Physics* **38**, pp. 181–203, 1997. Cited on page 1084.

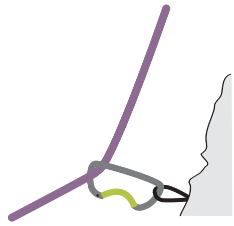
1164 The atomic masses, as given by IUPAC, can be found in *Pure and Applied Chemistry* **73**, pp. 667–683, 2001, or on the http://www.iupac.org website. For an isotope mass list, see the http://csnwww.in2p3.fr website. Cited on pages 1086 and 1095.

1165 The metallic, covalent and Van-der-Waals radii are from NATHANIEL W. ALCOCK, *Bonding and Structure*, Ellis Horwood 1999. The text also explains in detail how the radii are defined and measured. Cited on page 1095.

1166 Almost everything known about each element and its chemistry can be found in the encyclopedic GMELIN, *Handbuch der anorganischen Chemie*, published from 1817 onwards. It presently has over 500 volumes, now all published in English under the title *Handbook of Inorganic and Organometallic Chemistry*, and contains at least one volume dedicated to each chemical element. On the same topic, an incredibly expensive book with an equally bad layout is PER ENHAG, *Encyclopedia of the Elements*, Wiley-VCH, 2004. Cited on page 1084.

Appendix D

Numbers and Spaces



A mathematician is a machine that transforms coffee into theorems.

Paul Erdős (b. 1913 Budapest, d. 1996 Warsaw)

Mathematical concepts can all be constructed from 'sets' and 'relations.' The most important ones were presented in the first intermezzo. In the following a few more advanced concepts are presented as simply and vividly as possible,* for all those who want to smell the passion of mathematics. Ref. 1167

In particular, we shall expand the range of *algebraic* and the range of *topological* structures; the third basic type of mathematical structures, *order* structures, are not so important in physics. Mathematicians are not only concerned with the exploration of concepts, but always also with their *classification*. Whenever a new mathematical concept is introduced, mathematicians try to classify all the possible cases and types. Most spectacularly this has been achieved for the different types of numbers, of simple groups and for many types of spaces and manifolds.

Numbers as mathematical structures

Challenge 1490 ny

A person that can solve $x^2 - 92y^2 = 1$ in less than a year is a mathematician.

Brahmagupta, (b. 598 Sindh, d. 668) (implied: solve in *integers*)

To be most efficient, we start with a short introduction to the vocabulary. Any mathematical system with the same properties as the *natural* numbers is called a *semi-ring*. Any mathematical system with the same properties as the *integers* is called a *ring*. (The term is due to David Hilbert.) More precisely, a *ring* $(R, +, \cdot)$ is a set R of elements with two binary operations, called *addition* and *multiplication*, usually written $+$ and $\cdot$ (or simply dropped), for which the following properties hold for all elements $a, b, c \in R$:

- R is a commutative group with respect to addition, i.e. $a+b \in R,\ a+b = b+a,\ a+0 = a,\ a+(-a) = a-a = 0$ as well as $a+(b+c) = (a+b)+c$
- R is closed under multiplication, i.e. $ab \in R$
- multiplication is associative, i.e. $a(bc) = (ab)c$
- distributivity holds, i.e. $a(b+c) = ab + ac$ and $(b+c)a = ba + ca$.

* The opposite approach is taken by the delightful text by Carl E. Linderholm, *Mathematics Made Difficult*, Wolfe Publishing, 1971.

Defining properties such as these are also called *axioms*. Note that axioms are not basic beliefs, as often, but wrongly stated; axioms are the basic properties used in the definitions of a concept, in this case, that of ring.

A *semi-ring* is thus a set with all the properties of a ring, except that the existence of neutral and negative elements for addition is replaced by the weaker requirement that if $a + c = b + c$ then $a = b$.

To incorporate division and define the rational numbers, we need an additional concept. A *field* K is a ring with

- an identity 1, such that all elements a obey $1a = a$,
- at least one element different from zero, and most importantly
- a (multiplicative) inverse a^{-1} for every element $a \neq 0$.

A ring or field are said to be *commutative* if the multiplication is commutative. A non-commutative field is also called a *skew field*. Fields can be finite or infinite. (A field or a ring is characterized by its *characteristic* p. It is the smallest number of times one has to add 1 to itself to give zero. If there is no such number the characteristic is set to 0. p is always a prime number.) All finite fields are commutative. In a field, all equations of the type $cx = b$ and $xc = b$ $(c \neq 0)$ have solutions for x; they are unique if $b \neq 0$. To sum up sloppily by focusing on the most important property, a field is a set of elements for which, together with addition, subtraction and multiplication, a *division* (by non-zero elements) is also defined. The *rational numbers* are the simplest field that incorporates the integers.

The system of the *real numbers* is the minimal extension of the rationals which is continuous and totally ordered.*

However, the concept of 'number' is not limited to these examples. ** The simplest generalization is achieved by extending numbers to manifolds of more than one dimension.

Complex numbers

Complex numbers are defined by $z = a + ib$. The generators of the complex numbers, 1 and i, obey the well known algebra

$$\begin{array}{c|cc} \cdot & 1 & i \\ \hline 1 & 1 & i \\ i & i & -1 \end{array} \tag{793}$$

often written as $i = +\sqrt{-1}$.

The *complex conjugate* z^*, also written $\bar{z}$, of a complex number $z = a + ib$ is defined as $z^* = a - ib$. The *absolute value* $|z|$ of a complex number is defined as $|z| = \sqrt{zz^*} = \sqrt{z^*z} = \sqrt{a^2 + b^2}$. It defines a *norm* on the vector space of the complex numbers. From

* A set is *totally ordered* if there exists a binary relation that allows to say about any two elements who is the predecessor of the other in a consistent way.

** An excellent introduction into number systems in mathematics is the book H.-D. Ebbinghaus & al., *Zahlen*, 3. Auflage, Springer Verlag 1993. It is also available in English, under the title *Numbers*, Springer Verlag, 1990.

$|wz| = |w|\,|z|$ follows the *two-squares theorem*

$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) = (a_1 b_1 - a_2 b_2)^2 + (a_1 b_2 + a_2 b_1)^2 \qquad (794)$$

valid for all real numbers a_i, b_i. It was already known, in its version for integers, to Diophantus of Alexandria.

This means that complex numbers can also be written as a a couple (a, A), with their addition defined as $(a, A) + (b, B) = (a + b, A + B)$ and their multiplication defined as $(a, A) \cdot (b, B) = (ab - AB, aB + bA)$. The two component writing allows to identify complex numbers with the points on a plane. Therefore, translating the definition of multiplication into geometrical language allows to rapidly prove geometrical theorems, such as the one of Figure 375.

Challenge 1491 e

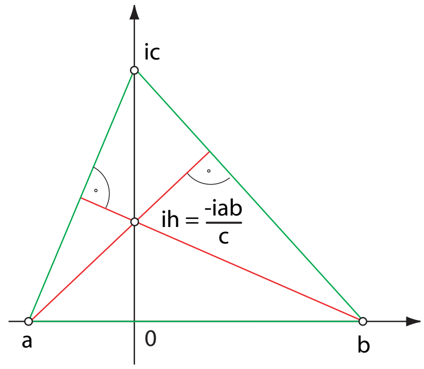


Figure 375 A property of triangles easily provable with complex numbers

Complex numbers can also be represented as 2×2 matrices of the form

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \quad \text{with} \quad a, b \in \mathbf{R} \,. \qquad (795)$$

Usual matrix addition and multiplication then give the same result as complex addition and multiplication. In this way, complex numbers can be represented by a special type of real matrices. What is $|z|$ in matrix language?

Challenge 1492 n

The set $\mathbf{C}$ of complex numbers with the mentioned addition and multiplication forms a commutative two-dimensional vector space. In the field of complex numbers, quadratic equations $az^2 + bz + c = 0$ for an unknown z always have two solutions.

Page 1107

Challenge 1493 ny

Complex numbers can be used to describe the position of the points of a plane. Rotations around the origin can be described by multiplications by a complex number of unit length. Since complex numbers describe the two-dimensional plane, any two-dimensional quantity can be described with them. That is why electrical engineers use complex numbers to describe quantities with phases, such as alternating currents or electrical fields in space.

Writing complex numbers of unit length as $\cos\theta + i\sin\theta$ is useful for remembering angle addition formulae. Since one has $\cos n\theta + i\sin n\theta = (\cos\theta + i\sin\theta)^n$, one can follow $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\sin\theta\cos\theta$.

Challenge 1494 e

By the way, there are as many complex numbers as there are real numbers. Are you able to show this?

Challenge 1495 n

Love is complex: it has real
and imaginary parts.

Quaternions

The position of the points on a line can be described by real numbers. Complex numbers can be used to describe the position of the points of a plane. If one tries to generalize the idea of a number to higher dimensional spaces, it turns out that no number system can be defined for *three*-dimensional space. However a new number system, the *quaternions*, can be constructed from the points of *four*-dimensional space, but only if the requirement of commutativity of multiplication is dropped. In fact, no number system can be defined for dimensions other than 1, 2 and 4. The quaternions were discovered by several mathematicians in the nineteenth century, among them Hamilton,* who studied them for a long part of his life. In fact, Maxwell's electrodynamics was formulated with quaternions before it was with three-dimensional vectors.

Ref. 1169

Page 1109

The quaternions **H** form a 4-dimensional algebra over the reals with the basis 1, i, j, k satisfying

$$\begin{array}{c|cccc} \cdot & 1 & i & j & k \\ \hline 1 & 1 & i & j & k \\ i & i & -1 & k & -j \\ j & j & -k & -1 & i \\ k & k & j & -i & -1 \end{array} \qquad (796)$$

which is also often written $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$. The quaternions 1, i, j, k are also called *basic units* or *generators*. The missing symmetry along the diagonal of the table shows the lack of commutativity of quaternionic multiplication. With the quaternions, the idea of a non-commutative product appeared for the first time in mathematics. Despite this restriction, the multiplication of quaternions remains associative. As a consequence of non-commutativity, polynomial equations in quaternions have many more solutions than in complex numbers; just search for all solutions of the equation $X^2 + 1 = 0$ to find out.

Challenge 1496 ny

Every quaternion X can be written in the form

$$X = x_0 + x_1 i + x_2 j + x_3 k = x_0 + \mathbf{v} = (x_0, x_1, x_2, x_3) = (x_0, \mathbf{v}) , \qquad (797)$$

where x_0 is called the *scalar* part and $\mathbf{v}$ the *vector* part. The multiplication is thus defined as $(x, \mathbf{v})(y, \mathbf{w}) = (xy - \mathbf{v}\cdot\mathbf{w}, x\mathbf{w} + y\mathbf{v} + \mathbf{v}\times\mathbf{w})$. The multiplication of two general quaternions can be written as

$$\begin{aligned}(a_1, b_1, c_1, d_1)(a_2, b_2, c_2, d_2) = (a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2\ ,\ a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2,\\ a_1c_2 - b_1d_2 + c_1a_2 + d_1b_2, a_1d_2 + b_1c_2 - c_1b_2 + d_1a_2)\ .\end{aligned} \qquad (798)$$

The conjugate quaternion $\overline{X}$ is defined as $\overline{X} = x_0 - \mathbf{v}$, so that $\overline{XY} = \overline{Y}\,\overline{X}$. The *norm* $|X|$ of a quaternion X is defined as $|X|^2 = X\overline{X} = \overline{X}X = x_0^2 + x_1^2 + x_2^2 + x_3^2 = x_0^2 + \mathbf{v}^2$. The norm is multiplicative, i.e. $|XY| = |X|\,|Y|$.

* William Rowan Hamilton (b. 1805 Dublin, d. 1865 Dunsink), Irish child prodigy, mathematician, named the quaternions after an expression from the vulgate (act. apost. 12, 4).

Unlike complex numbers, every quaternion is related to its complex conjugate by

$$\overline{X} = -\tfrac{1}{2}(X + iXi + jXj + kXk) . \tag{799}$$

In the case of complex numbers, a number and its conjugate are independent variables; for quaternions, this is not the case. As a result, functions of quaternions have less descriptive power than functions of complex variables.

The relation $|XY| = |X|\,|Y|$ implies the *four-squares theorem*

$$\begin{aligned} &(a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) \\ = &(a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4)^2 + (a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3)^2 \\ &+(a_1b_3 + a_3b_1 + a_4b_2 - a_2b_4)^2 + (a_1b_4 + a_4b_1 + a_2b_3 - a_3b_2)^2 \end{aligned} \tag{800}$$

valid for all real numbers a_i and b_i, and thus also for any set of eight integers. It was discovered in 1748 by Leonhard Euler (1707–1783) when trying to prove that each integer is the sum of four squares. (That proof was found only in 1770, by Joseph Lagrange.)

Hamilton thought that a quaternion with zero scalar part, which he simply called a *vector* – a term which he invented –, could be identified with an ordinary 3-dimensional translation vector; but this is wrong. Therefore, such a quaternion is now called a *pure*, or a *homogeneous*, or again, an *imaginary* quaternion. The product of two pure quaternions $V = (0, \mathbf{v})$ and $W = (0, \mathbf{w})$ is given by $VW = (-\mathbf{v}\cdot\mathbf{w}, \mathbf{v}\times\mathbf{w})$, where $\cdot$ denotes the scalar product and $\times$ denotes the vector product. Note that any general quaternion can be written as the ratio of two pure quaternions.

In reality, a pure quaternion $(0, \mathbf{v})$ does not behave under coordinate transformations like a (modern) vector; in fact, a pure quaternion represents a rotation by the angle π or 180°C around the axis defined by the direction $\mathbf{v} = (v_x, v_y, v_z)$.

Challenge 1497 ny

Challenge 1498 n

Can you find out what –1 is? In other words, quaternions can be used to describe the belt trick, if the multiplication VW of two quaternions is taken to mean that rotation V is performed *after* rotation W.

Page 730

It turns out that in three-dimensional space, a *general* rotation about the origin can be described by a *unit* quaternion, also called a *normed* quaternion, for which $|Q| = 1$. Such a quaternion can be written as $(\cos\theta/2, \mathbf{n}\sin\theta/2)$, where $\mathbf{n} = (n_x, n_y, n_z)$ is the normed vector describing the direction of the rotation axis and θ is the rotation angle. Such a unit quaternion $Q = (\cos\theta/2, \mathbf{n}\sin\theta/2)$ rotates a pure quaternion $V = (0, \mathbf{v})$ into another pure quaternion $W = (0, \mathbf{w})$ given by

$$W = QVQ^* . \tag{801}$$

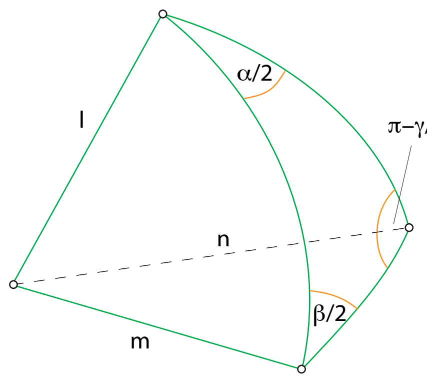


Figure 376 Combinations of rotations

In this case, when using pure quaternions such as V or W to describe positions, unit quaternions can be used to describe rotations and to calculate coordinate changes. The concatenation of two rotations is then given as the product of the corresponding unit quaternions. Indeed, a rotation by an angle α about the axis $\mathbf{l}$ followed by a rotation by an angle β about the axis $\mathbf{m}$ gives a rotation by an angle γ about axis $\mathbf{n}$, with the values determined by

$$(\cos\gamma/2, \sin\gamma/2\mathbf{n}) = (\cos\alpha/2, \sin\alpha/2\mathbf{l})(\cos\beta/2, \sin\beta/2\mathbf{m}) \ . \qquad (802)$$

One way to show the result graphically is given in Figure 376. By drawing a triangle on a unit sphere and taking care to remember the factor 1/2 in the angles, the combination of two rotations can simply be determined by drawing a triangle on the surface of the sphere.

The interpretation of quaternions as rotations by half angles is also illustrated, in a somewhat different way, in the motion of any hand and arm. Keeping the left arm straight, defining its three possible 90 degree motions as i, j and k, and taking concatenation as multiplication, the motion of our arm follows the same 'laws' as those of pure unit quaternions. In this analogy, the rotation angle of the arm is half the rotation angle of the corresponding quaternion. Ref. 1168

The reason for the half-angle behaviour of rotations is their non-commutativity under composition. It can be specified more precisely using mathematical language. The rotations in 3 dimensions around a point form the Special Orthogonal group in 3 dimensions, in short SO(3). But the motions of a hand attached to a shoulder via an arm form another group, isomorphic to the Lie group SU(2). The difference is due to the appearance of half angles in the parametrization of rotations; indeed, the above parametrizations imply that a rotation by 2π corresponds to a multiplication by -1. Only in the twentieth century it was realized that fundamental physical observables behaving like hands attached to arms do exist: they are called *spinors*. More on spinors can be found in the section on permutation symmetry, where belts are used as well as arms. In short, the group SU(2) of the quaternions is the *double cover* of the rotation group SO(3). Page 1121 Page 730 Ref. 1170

The simplicity of the description of rotations and positions with quaternions is used in robotics, in astronomy and in flight simulators; this is due to the especially simple coding of coordinate transformations quaternions provide. Also in three-dimensional graphic visualization software, quaternions are often used to calculate the path taken by repeatedly reflected light rays.

The algebra of the quaternions is the unique associative noncommutative finite-dimensional normed algebra with an identity over the field of real numbers. Quaternions form a noncommutative field, i.e. a skew field, in which the inverse of a quaternion X is $\overline{X}/N(X)$. This allows to define a division of quaternions. Therefore quaternions are said to form a *division algebra*. In fact, the quaternions $\mathbf{H}$, the complex numbers $\mathbf{C}$ and the reals $\mathbf{R}$ form the only three examples of finite dimensional associative division algebras. In other words, the skew-field of quaternions is the unique finite-dimensional real associative non-commutative algebra without divisors of zero. The *centre* of the quaternions, i.e. the set of those quaternions commuting with all quaternions, are the reals.

Like the complex numbers, quaternions can be represented as matrices of the form

$$\begin{pmatrix} A & B \\ -B^* & A^* \end{pmatrix} \text{with } A, B \in \mathbf{C}, \quad \text{or as} \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix} \text{with } a, b, c, d \in \mathbf{R}, \quad (803)$$

where $A = a + ib$, $B = c + id$ and the quaternion X is $X = A + Bj = a + ib + jc + kd$; usual matrix addition and multiplication then give the same result as quaternionic addition and multiplication.

The generators of the quaternions can be realized for example as

$$1 : \sigma_0 \quad , \quad i : -i\sigma_1 \quad , \quad j : -i\sigma_2 \quad , \quad k : -i\sigma_3 \qquad (804)$$

where the σ_n are the Pauli spin matrices.*

Real 4×4 representations are not unique, as

$$\begin{pmatrix} a & b & -d & -c \\ -b & a & -c & d \\ d & c & a & b \\ c & -d & -b & a \end{pmatrix} \qquad (806)$$

Challenge 1499 ny

shows; however, no representation by 3×3 matrices is possible.

These matrices contain real and complex elements, which pose no special problems. In contrast, when matrices with quaternionic elements are constructed, care has to be taken, because simple relations, such as $\mathrm{tr}AB = \mathrm{tr}BA$ are *not* fulfilled in general, since quaternionic multiplication is not commutative.

What do we learn from quaternions for the description of nature? First of all, binary rotations are similar to positions and thus to translations. Are rotations the basic operations? Is it possible that translations are only shadows of rotations? The way that translations are connected to rotations is investigated in the third part of the mountain ascent.

When Maxwell wrote down his equations of electrodynamics, he used quaternion notation. The now usual 3-vector notation was introduced later by other scientists, notably by Hertz and Heaviside. Maxwell's equations of electrodynamics, can be written in various ways using quaternions. The simplest way is found when one keeps a distinction between $\sqrt{-1}$ and the units i, j, k of the quaternions. One then can write all of electrody-

Page 503

Ref. 1169

* The *Pauli spin matrices* are the complex, Hermitean matrices

$$\sigma_0 = \mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad , \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad (805)$$

whose eigenvalues are ± 1; they satisfy the relations $[\sigma_i, \sigma_k]_+ = 2\,\delta_{ik}$ and $[\sigma_i, \sigma_k] = 2i\,\varepsilon_{ikl}\,\sigma_l$. The linear combinations $\sigma_\pm = \frac{1}{2}(\sigma_1 \pm \sigma_2)$ are also frequently used. By the way, another possible representation of the quaternions is $i : i\sigma_3, j : i\sigma_2, k : i\sigma_1$.

Challenge 1500 n

namics in a single equation:

$$dF = -\frac{Q}{\varepsilon_0} \tag{807}$$

where F is the generalized electromagnetic field Q the generalized charge. They are defined by

$$\begin{aligned} F &= E + \sqrt{-1}\,cB \\ E &= iE_x + jE_y + kE_z \\ B &= iB_x + jB_y + kB_z \\ d &= \delta + \sqrt{-1}\,\partial_t/c \\ \delta &= i\partial_x + j\partial_y + k\partial_z \\ Q &= \rho + \sqrt{-1}\,J/c \end{aligned} \tag{808}$$

where the fields E and B and the charge distributions ρ and J have the usual meaning. The content of equation 807 for the electromagnetic field is exactly the same as the usual formulation.

Despite their charm, quaternions do not seem too useful for the reformulation special relativity; the main reason is the wrong sign in the expression for their norm. Therefore, relativity and space-time are usually described with real numbers.

Octonions

In the same way that the quaternions are constructed from complex numbers, octonions can be constructed from quaternions, as done by Arthur Cayley (1821–1895). *Octonions* or *octaves* are the elements of an 8-dimensional algebra over the reals with the generators 1, i_n with $n = 1...7$ satisfying

(809)

	1	i_1	i_2	i_3	i_4	i_5	i_6	i_7
1	1	i_1	i_2	i_3	i_4	i_5	i_6	i_7
i_1	i_1	-1	i_3	$-i_2$	i_5	$-i_4$	i_7	$-i_6$
i_2	i_2	$-i_3$	-1	i_1	$-i_6$	i_7	i_4	$-i_5$
i_3	i_3	i_2	$-i_1$	-1	i_7	i_6	$-i_5$	$-i_4$
i_4	i_4	$-i_5$	i_6	$-i_7$	-1	i_1	$-i_2$	i_3
i_5	i_5	i_4	$-i_7$	$-i_6$	$-i_1$	-1	i_3	i_2
i_6	i_6	$-i_7$	$-i_4$	i_5	i_2	$-i_3$	-1	i_1
i_7	i_7	i_6	i_5	i_4	$-i_3$	$-i_2$	$-i_1$	-1

Nineteen other, equivalent multiplication tables are also possible. This algebra is called the *Cayley algebra*; it has an identity and a unique division. The algebra is non-commutative and also non-associative. It is however, *alternative*, meaning that for all elements, one has $x(xy) = x^2y$ and $(xy)y = xy^2$, a property somewhat weaker than associativity. It is the only 8-dimensional real alternative algebra without zero divisors. For this last reason, the set $\mathbf{\Omega}$ of all octonions does not form a field nor a ring, so that the old designation of

'Cayley numbers' has been abandoned. The octonions are the most general hypercomplex 'numbers' whose norm is multiplicative. Associativity is not satisfied, since $(i_n i_m) i_l = \pm i_n (i_m i_l)$, where the minus sign is valid for combination of indices which belong to those triads, such as 1-2-4, which are not quaternionic.

Octonions can be represented as matrices of the form

$$\begin{pmatrix} A & B \\ -\bar{B} & \bar{A} \end{pmatrix} \quad \text{where} \quad A, B \in \mathbf{H} \quad , \text{ or as real } 8 \times 8 \text{ matrices.} \tag{810}$$

Matrix multiplication then gives the same result as octonionic multiplication.

The relation $|wz| = |w|\,|z|$ allows to deduce the impressive *eight-squares theorem*

$$\begin{aligned}
&(a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2 + a_6^2 + a_7^2 + a_8^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2 + b_5^2 + b_6^2 + b_7^2 + b_8^2) \\
&\quad = (a_1 b_1 - a_2 b_2 - a_3 b_3 - a_4 b_4 - a_5 b_5 - a_6 b_6 - a_7 b_7 - a_8 b_8)^2 \\
&\quad + (a_1 b_2 + a_2 b_1 + a_3 b_4 - a_4 b_3 + a_5 b_6 - a_6 b_5 - a_7 b_8 + a_8 b_7)^2 \\
&\quad + (a_1 b_3 - a_2 b_4 + a_3 b_1 + a_4 b_2 + a_5 b_7 + a_6 b_8 - a_7 b_5 - a_8 b_6)^2 \\
&\quad + (a_1 b_4 + a_2 b_3 - a_3 b_2 + a_4 b_1 + a_5 b_8 - a_6 b_7 + a_7 b_6 - a_8 b_5)^2 \\
&\quad + (a_1 b_5 - a_2 b_6 - a_3 b_7 - a_4 b_8 + a_5 b_1 + a_6 b_2 + a_7 b_3 + a_8 b_4)^2 \\
&\quad + (a_1 b_6 + a_2 b_5 - a_3 b_8 + a_4 b_7 - a_5 b_2 + a_6 b_1 - a_7 b_4 + a_8 b_3)^2 \\
&\quad + (a_1 b_7 + a_2 b_8 + a_3 b_5 - a_4 b_6 - a_5 b_3 + a_6 b_4 + a_7 b_1 - a_8 b_2)^2 \\
&\quad + (a_1 b_8 - a_2 b_7 + a_3 b_6 + a_4 b_5 - a_5 b_4 - a_6 b_3 + a_7 b_2 + a_8 b_1)^2
\end{aligned} \tag{811}$$

valid for all real numbers a_i and b_i and thus in particular also for all integers. It was discovered in 1818 by Carl Friedrich Degen (1766–1825), and then rediscovered in 1844 by John Graves and in 1845 by Cayley. There is no generalization to higher numbers of squares, a fact proven by Adolf Hurwitz (1859–1919) in 1898.

As a note, the octonions can be used to show that a vector product is not only possible in dimensions 3. A *vector product* or *cross product* is an operation satisfying

$$\begin{aligned}
u \times v = -v \times u & \quad \textit{anticommutativity} \\
(u \times v) w = u (v \times w) & \quad \textit{exchange rule.}
\end{aligned} \tag{812}$$

Using the definition

$$X \times Y = \frac{1}{2}(XY - YX) \,, \tag{813}$$

the $\times$-products of imaginary quaternions, i.e. of quaternions of the sort $(0, \mathbf{u})$, are again imaginary, and the $\mathbf{u}$'s obey the usual vector product, thus fulfilling (812). Interestingly, using definition (813) for *octonions* is possible. In that case the product of imaginary octonions yields only imaginary octonions, and the $\mathbf{u}$'s also follow expression (812). In fact, this is the only other non-trivial example possible. Thus a vector product exists only in 3 and in 7 dimensions.

Ref. 1171

Challenge 1501 e

Other types of numbers

The process of construction of a new system of hypercomplex 'numbers' or real algebras by 'doubling' a given one can be continued ad infinitum. However, octonions, *sedenions* and all the following doublings are neither rings nor fields, but only non-associative algebras with unity. Other finite-dimensional algebras with unit element over the reals, once generally called hypercomplex 'numbers', can also be defined, such as 'dual numbers', 'double numbers', 'Clifford–Lifshitz numbers' etc. They play no special role in physics.

Mathematicians also have defined number fields which have 'one and a half' dimensions, such as algebraic number fields. There is also a generalization of the concept of integers to the complex domain, the *Gaussian integers*, defined as $n + im$. Gauß even defined what now are known as *Gaussian primes*. (Can you find out how?) They are not used in the description of nature, but are important in number theory.

Ref. 1172
Challenge 1502 n

In the old days physicists used to call quantum mechanical operators 'q-numbers.' But this term has now fallen out of fashion.

Other extensions of the natural numbers are those which include numbers larger than the smallest type of infinity. The most important *transfinite numbers* are the *ordinals*, the *cardinals* and the mentioned *surreals*. The ordinals are essentially the infinite integers (and the finite ones), whereas the surreals are the infinite (and finite) reals. The surreals were defined in the first intermezzo. They are to the ordinal numbers what the reals are to the integers: they fill up all the gaps in between. Interestingly, for the surreals, the summation of many divergent series in **R** converge. Can you find one example?

Ref. 1173
Page 606
Challenge 1503 ny

The surreals also include infinitely small numbers. That is also the case for the numbers of *nonstandard analysis*, also called *hyperreals*. In both number systems, in contrast to the case of the real numbers, the numbers $0.999\overline{9}$ and 1 do not coincide, but are separated by infinitely many other numbers.

Ref. 1171

Grassmann numbers

With the discovery of supersymmetry, another type of numbers became important, the *Grassmann numbers*.* They are in fact a special type of hypercomplex 'numbers'. In supersymmetric Lagrangians, fields depend on two types of coordinates: on the usual real space-time coordinates and additionally on Grassmann coordinates.

Grassmann numbers, also called *fermionic coordinates*, θ have the defining properties

$$\theta^2 = 0 \quad \text{and} \quad \theta_i\theta_j + \theta_j\theta_i = 0 \; . \tag{814}$$

Challenge 1504 ny

You may want to look for a representation of these numbers. More about their use can be found in the section on supersymmetry.

Vector spaces

Vector spaces, also called linear spaces, are mathematical generalizations of certain aspects of the intuitive three-dimensional space. Any set of elements that can be added

* Hermann Günther Grassmann (1809–1877), school teacher in Stettin, and one of the most profound mathematicians of the nineteenth century.

together and also be multiplied by numbers is called a vector space, if the result is again in the set and the usual rules of calculation hold.

More precisely, a *vector space* over a number field K is a set of elements, called *vectors* in this case, for which a vector addition and a *scalar multiplication* is defined for all vectors a, b, c and for all numbers s and r from K with the properties

$$\begin{aligned}(a+b)+c = a+(b+c) = a+b+c & \quad \textit{associativity of vector addition}\\ n+a = a & \quad \textit{existence of null vector}\\ (-a)+a = n & \quad \textit{existence of negative vector}\\ 1a = a & \quad \textit{regularity of scalar multiplication}\\ (s+r)(a+b) = sa+sb+ra+rb & \quad \textit{complete distributivity of scalar multiplication}\end{aligned} \tag{815}$$

If the field K, whose elements are called *scalars* in this context, is taken to be the real (complex, quaternionic) numbers, one speaks of a real (complex, quaternionic) vector space. Vector spaces are also called *linear vector spaces* or simply *linear spaces*.

The complex numbers, the set of all functions defined on the real line, the set of all polynomials, the set of matrices of given number of rows and columns all form vector spaces. In mathematics, a vector is thus a more general concept than in physics. Physical vectors are more specialized objects, namely elements of normed inner product spaces. To define them one first needs the concept of metric space.

A *metric space* is a vector space with a metric, i.e. a way to define distances between elements. A relation $d(a,b)$ between elements is called a metric if

$$\begin{aligned}d(a,b) \geqslant 0 & \quad \textit{positivity of metric}\\ d(a,b)+d(b,c) \geqslant d(a,c) & \quad \textit{triangle inequality}\\ d(a,b)=0 \quad \text{if and only if} \quad a=b & \quad \textit{regularity of metric}\end{aligned} \tag{816}$$

Challenge 1505 n For example, measuring the distance between cities in France, i.e. points on a surface, by the shortest distance of travel via Paris, except in the case if they both lie on a line already going through Paris, defines a metric between the points in France.

A *normed* vector space is, obviously, a linear space with norm, or 'length' of a vector. A norm is a positive (or vanishing) number $||a||$ defined for each vector a with the properties

$$\begin{aligned}||ra|| = |r|\,||a|| & \quad \textit{linearity of norm}\\ ||a+b|| \leqslant ||a||+||b|| & \quad \textit{triangle inequality}\\ ||a||=0 \quad \text{only if} \quad a=0 & \quad \textit{regularity}\end{aligned} \tag{817}$$

Challenge 1506 ny Usually there are many ways to define a norm for a given space. Note that a norm can always be used to define a metric by setting

$$d(a,b) = ||a-b|| \tag{818}$$

so that all normed spaces are also metric spaces. This norm is the standard choice. It is often defined with the help of an inner product.

The most special linear spaces are *inner product spaces*. They are vector spaces with an *inner product*, also called *scalar product* · (not to be confused with the scalar multiplication!). For an inner product in the *real* case the properties of

$$\begin{aligned} a \cdot b &= b \cdot a && \textit{commutativity of scalar product} \\ (ra)\cdot(sb) &= rs(a \cdot b) && \textit{bilinearity of scalar product} \\ (a+b)\cdot c &= a \cdot c + b \cdot d && \textit{left distributivity of scalar product} \\ a\cdot(b+c) &= a\cdot b + a \cdot c && \textit{right distributivity of scalar product} \\ a \cdot a &\geqslant 0 && \textit{positivity of scalar product} \\ a\cdot a = 0 \quad \text{if and only if} \quad a &= 0 && \textit{regularity of scalar product} \end{aligned} \tag{819}$$

hold for all vectors a, b and all scalars r, s. A *real* inner product space (of finite dimension) is also called a *Euclidean* vector space. The set of all velocities, the set of all positions, or the set of all possible momenta form such spaces.

In the *complex* case this definition is extended to*

$$\begin{aligned} ab &= \overline{(ba)} = \overline{b}\overline{a} && \textit{Hermitean property} \\ (ra)(sb) &= \overline{r}s(ab) && \textit{sesquilinearity of scalar product} \\ (a+b)c &= ac + bd && \textit{left distributivity of scalar product} \\ a(b+c) &= ab + ac && \textit{right distributivity of scalar product} \\ aa &\geqslant 0 && \textit{positivity of scalar product} \\ aa = 0 \quad \text{if and only if} \quad a &= 0 && \textit{regularity of scalar product} \end{aligned} \tag{820}$$

hold for all vectors a, b and all scalars r, s. A *complex* inner product space (of finite dimension) is also called a *unitary* or *Hermitean* vector space. If the inner product space is *complete*, it is called, especially in the infinite-dimensional complex case, a *Hilbert space*. The space of all possible states of a quantum system form a Hilbert space.

All inner product spaces are also metric spaces and thus normed spaces, if the metric is defined, as usually done, by

$$d(a,b) = \sqrt{(a-b)(a-b)}\ . \tag{821}$$

Inner product spaces allow to speak about the length and the direction of vectors, as we are used to in physics. In addition, inner product spaces allow to define a *basis*, the mathematical concept used to define coordinate systems.

Challenge 1507 ny What is the simplest possible vector space?

The *dimension* of a vector space is the number of linearly independent basis vectors.

Challenge 1508 ny Can you define theses terms in detail?

Challenge 1509 ny Which vector spaces are of importance in physics?

Algebras

The term *algebra* is used in mathematics with three different, but loosely related meanings. It denotes a part of mathematics, as in 'I hated algebra at school'; it further denotes in general any formal rules that are obeyed by abstract objects, as e.g. in the expression

* The term *sesquilinear* is Latin for 'one-and-a-half-linear'. Sometimes however, the half-linearity is assumed in the other argument.

'tensor algebra'. Finally – and that is the only meaning used here – an algebra denotes a *specific* mathematical structure.

The most intuitive definition focuses on the central aspect: an algebra is a set of vectors that can also be multiplied to give another vector. More precisely, an *algebra* is a vector space that is also a ring. (The concept is due to Benjamin Peirce (1809–1880), father of Charles Sanders Peirce.) A ring is – like the integers – a set for which an addition and a multiplication is defined. This implies that in an algebra, there are *three* types of multiplications:

Page 1098

- scalar multiplication: the product of two algebra elements (vectors) x and y is a scalar $c = x \cdot y$;
- s-multiplication: the numerical multiple of a vector x with a scalar c is another vector $y = cx$;
- the (main) algebraic multiplication: the product of two algebra elements x and y is another algebra element $z = xy$.

The precise definition of algebra thus only needs to define the precise properties of the (main) multiplication and to specify the number field K. All algebras have

$$x(y+z) = xy + xz \quad , \quad (x+y)z = xz + yz \quad \textit{distributivity of multiplication}$$
$$c(xy) = (cx)y = x(cy) \quad \textit{bilinearity} \qquad (822)$$

for all scalars $c \in$ K. For example, the set of all linear transformations in an n-dimensional linear space, such as the translations on a plane, in space or in time, are linear algebras. So is the set of observables of a quantum mechanical system.*

An *associative algebra* is an algebra whose multiplication has the additional property that

$$x(yz) = (xy)z \quad \textit{associativity} \qquad (824)$$

Most physical algebras are linear and associative. Therefore, a linear associative algebra is often simply called an associative algebra or, even shorter, an algebra.**

The set of multiples of the unit 1 of the algebra is called the *field of scalars* scal(A) of

* *Linear transformations* are mathematical objects which transform a vector into another with the property that sums and multiples of vectors are transformed into sums and the multiples of the transformed vectors. Are you able to give the set of all possible linear transformations of points on a plane? And in space? And in Minkowski space?

Challenge 1510 n

All linear transformations transform some special vectors, called *eigenvectors* – from the German word 'eigen' meaning 'self' – into multiples of themselves. In other words, if a transformation T has the effect to yield

$$Te = \lambda e \ , \qquad (823)$$

then the vector e is called an *eigenvector* and λ the associated *eigenvalue* of the transformation T. The set of all eigenvalues of a transformation T is called the *spectrum* of T. Physicists did not care for these mathematical concepts until they discovered quantum theory. Quantum theory showed that observables are transformations in Hilbert space, because any measurement interacts with a system and thus transforms it. Quantum mechanical experiments also showed that a measurement result for an observable can only be one of the eigenvalues – or one element of the spectrum – of the corresponding transformation. The state of the system after the measurement is given by the eigenvector of the measured eigenvalue. Therefore every expert on motion must know what an eigenvalue is.

Page 693
Page 750

** Note that a non-associative algebra does not necessarily possess a matrix representation.

the algebra A. The field of scalars is also a subalgebra of A. The field of scalars and the scalars themselves behave in the same way. We explore a few examples.

Challenge 1511 e All polynomials in one variable (or in several variables) form an algebra. It is commutative and infinite-dimensional. The constant polynomials form the field of scalars.

Challenge 1512 ny The set of $n \times n$ matrices, with the proper operations, also forms an algebra. It is n^2-dimensional. Those diagonal matrices – with all non-diagonal elements set to zero – whose diagonal elements have all the same value form the field of scalars. By the way, what is the matrix product and the scalar product of two matrices?

Challenge 1513 n The set of all real functions also forms an algebra. Can you specify the most straightforward multiplication? The constant functions form the field of scalars.

A *star algebra*, also written $*$-*algebra*, is an algebra over the *complex* numbers for which there is a mapping $* : A \to A,\ x \mapsto x^*$, called an *involution*, with the properties

$$\begin{aligned} (x^*)^* &= x \\ (x+y)^* &= x^* + y^* \\ (cx)^* &= \bar{c}x^* \quad \text{for all} \quad c \in \mathbb{C} \\ (xy)^* &= y^* x^* \end{aligned} \tag{825}$$

valid for all elements x, y of the algebra A. The element x^* is called the *adjoint* of x. Star algebras are the main structure used in quantum mechanics, since quantum mechanical observables form a $*$-algebra.

A C$*$-*algebra* is a Banach algebra, i.e., a complete normed algebra, over the complex numbers with an involution $*$, i.e. a function which is its own inverse, so that the norm $\|x\|$ of an element x can be defined as

$$\|x\|^2 = x^* \, x \, . \tag{826}$$

Challenge 1514 n Algebras are *complete* if Cauchy sequences converge. The name C comes from 'continuous functions'; the *bounded* continuous functions form such an algebra with a properly defined norm. Can you find it?

All C$*$-algebras contain a space of Hermitean elements (which have a real spectrum), a set of normal elements, a multiplicative group of unitary elements and a set of positive elements (with non-negative spectrum).

One important type of algebra used in mathematics deserves to be mentioned. A *division algebra* is an algebra for which the equations $ax = b$ and $ya = b$ are uniquely solvable in x or y for all b and all $a \neq 0$. Obviously, all type of numbers must be division algebras. Division algebras are thus one way to generalize the concept of a number. One of the important results of modern mathematics states that (finite-dimensional) division algebras can *only* have dimension 1, like the reals, dimension 2, like the complex numbers, dimension 4, like the quaternions, or dimension 8, like the octonions. There is thus no way to generalize the concept of 'number' to other dimensions.

And now some fun. Imagine a *ring* A which contains a number field K as a subring (or 'field of scalars'). If the ring multiplication is defined in such a way that a general ring element multiplied with an element of K is the same as the scalar multiplication, then A is a vector space and thus an algebra – *provided* that every element of K commutes with every element of A. (In other words, the subring or subfield K must be *central*.)

For example, the quaternions H are a 4-dimensional real division algebra, but although H is a 2-dimensional complex vector space, it is *not* a complex algebra, because i does not commute with j (one has $ij = -ji = k$). In fact, there are no finite dimensional complex division algebras, and the only finite dimensional real associative division algebras are **R**, Cand **H**.

Now, if you are not afraid of getting a headache, think about this remark: every K-algebra is also an algebra over its field of scalars. For this reason, some mathematicians prefer to define an (associative) K-algebra simply as a ring which contains K as a central subfield.

In physics, those algebras related to symmetries play the most important role. We study them next.

Lie algebras

A Lie algebra is special type of algebra and thus of vector space. Lie algebras are the most important type of non-associative algebras. A vector space L over the field **R** (or **C**) with an additional binary operation [,] called *Lie multiplication* or the *commutator*, is called a real (or complex) Lie algebra if this operation fulfils the properties

$$\begin{aligned} [X,Y] &= -[Y,X] && \textit{antisymmetry} \\ [aX+bY,Z] &= a[X,Z]+b[Y,Z] && \textit{(left-)linearity} \\ [X,[Y,Z]]+[Y,[Z,X]]+[Z,[X,Y]] &= 0 && \textit{Jacobi identity} \end{aligned} \qquad (827)$$

Challenge 1515 e

for all elements $X, Y, Z \in L$ and for all $a, b \in$ **R** (or **C**). (Lie algebras are named after Sophus Lie.) The first two conditions together imply bilinearity. A Lie algebra is called *commutative* if $[X, Y] = 0$ for all elements X and Y. The *dimension* of the Lie algebra is the dimension of the vector space. A subspace N of a Lie algebra L is called an *ideal* if $[L, N] \subset N$; any ideal is also a *subalgebra*. A *maximal ideal* M which satisfies $[L, M] = 0$ is called the *centre* of L.

A Lie algebra is called a *linear* Lie algebra if its elements are linear transformations of another vector space V, simply said, if they are 'matrices'. It turns out that every finite dimensional Lie algebra is isomorphic to a linear Lie algebra. Therefore, by picturing the elements of Lie algebras in terms of matrices all finite dimensional cases are covered.

Page 1125

The name 'Lie algebra' was chosen because the *generators*, i.e. the infinitesimal elements of every Lie group form a Lie algebra. Since all important symmetries in nature form Lie groups, Lie algebras appear very frequently in physics. In mathematics, Lie algebras arise frequently because from any associative finite dimensional algebra in which the symbol $\cdot$ stands for its multiplication, a Lie algebra appears when defining the *commutator* by

$$[X,Y] = X \cdot Y - Y \cdot X \quad ; \qquad (828)$$

this fact gave the commutator its name. Lie algebras are non-associative; the mentioned definition of the commutator shows how to build it from an associative algebra.

Since Lie algebras are vector spaces, the elements T_i of a *basis* of the Lie algebra always obey the relation:

$$[T_i, T_j] = \sum_k c^k_{ij} T_k \tag{829}$$

where the numbers c^k_{ij} are called the structure constants of the Lie algebra. They depend on the chosen basis. Structure constants determine the Lie algebra completely. For example, the algebra of the Lie group SU(2), with the three generators defined by $T_a = \sigma^a/2i$, where the σ^a are the Pauli spin matrices, has the structure constants $C_{abc} = \varepsilon_{abc}$.*

Page 1104

Classification of Lie algebras

All Lie algebras can be divided in finite-dimensional and infinite dimensional ones. Every *finite-dimensional* Lie algebra turns out to be the (semidirect) sum of a semisimple and a solvable Lie algebra.

A Lie algebra is called *solvable* if, well, if it is not semisimple. Solvable Lie algebras have not been classified completely up to now.They are not important in physics.

A *semisimple* Lie algebra is a Lie algebra which has no non-zero solvable ideal. Other equivalent definitions are possible, depending on your taste:

- a semisimple Lie algebra does not contain non-zero abelian ideals,
- its Killing-form is non-singular, i.e. non-degenerate,
- it splits into the direct sum of non-abelian simple ideals (this decomposition is unique)
- every finite-dimensional linear representation is completely reducible
- the one-dimensional cohomology of g with values in an arbitrary finite-dimensional g-module is trivial.

All finite-dimensional semisimple Lie algebras have been completely classified. Every semisimple Lie algebra decomposes uniquely into a direct sum of *simple* Lie algebras. Simple Lie algebras can be complex or real.

The simple finite-dimensional *complex* Lie algebras all belong to four infinite classes and to five exceptional cases. The infinite classes are also called *classical* and are A_n for

---

* In the same ways as groups, Lie algebras can be represented by matrices, i.e. by linear operators. Representations of Lie algebras are important in physics because many continuous symmetry groups are Lie groups.

The *adjoint representation* of a Lie algebra with basis $a_1...a_n$ is the set of matrices $\mathrm{ad}(a)$ defined for each element a by

$$[a, a_j] = \sum_c \mathrm{ad}(a)_{cj} a_c \; . \tag{830}$$

It implies that $\mathrm{ad}(a)_{jk} = c^k_{ij}$, where c^k_{ij} are the structure constants of the Lie algebra. For a real Lie algebra, all elements of $\mathrm{ad}(a)$ are real for all $a \in L$.

Note that for any Lie algebra, a scalar product can be defined by setting

$$X \cdot Y = \mathrm{Tr}(\, \mathrm{ad}X \cdot \mathrm{ad}Y \,) \; . \tag{831}$$

This scalar product is symmetric and bilinear. The corresponding bilinear form is also called the *Killing form*, after the German mathematician Wilhelm Killing (1847–1923), the discoverer of the exceptional Lie groups. The Killing form is invariant under the action of any automorphism of the algebra L. In a given basis, one has

$$X \cdot Y = \mathrm{Tr}(\, (\mathrm{ad}X)^i_k \cdot (\mathrm{ad}Y)^s_i \,) = c^i_{lk} c^k_{si} x^l y^s = g_{ls} x^l y^s \tag{832}$$

where $g_{ls} = c^i_{lk} c^k_{si}$ is called the *Cartan metric tensor* of the Lie algebra L.

$n \geqslant 1$, corresponding to the Lie groups $SL(n)$ and their compact 'cousins' $SU(n)$, B_n for $n \geqslant 2$, corresponding to the Lie groups $SO(2n+1)$, C_n for $n \geqslant 3$, corresponding to the Lie groups $Sp(2n)$, and D_n for $n \geqslant 4$, corresponding to the Lie groups $SO(2n)$. These simple Lie algebras are defined as follows. A_n is the algebra of all skew-Hermitean $n \times n$ matrices, B_n and C_n are the algebras of the symmetric $n \times n$ matrices, and D_n is the algebra of the traceless $n \times n$ matrices.

The exceptional Lie algebras are G_2, F_4, E_6, E_7, E_8. In all cases, the index gives the number of roots. The dimension of the algebras is $A_n : n(n+2)$, B_n and $C_n : n(2n+1)$, $D_n : n(2n-1)$, $G_2 : 14$, $F_4 : 32$, $E_6 : 78$, $E_7 : 133$, $E_8 : 248$.

The simple and finite-dimensional *real* Lie algebras are more numerous; they follow from the list of complex Lie algebras. Moreover, for each complex Lie group, there is always one compact real one. Real Lie algebras are not so important in fundamental physics. Superalgebras play a role in systems with supersymmetry.

Ref. 1174

Of the large number of *infinite dimensional* Lie algebras only few are important in physics, among them the *Poincaré algebra*, the *Cartan algebra*, the *Virasoro algebra* and a few other Kac–Moody algebras.

Lie superalgebras

Lie superalgebras arise when the concept of Lie algebra is extended to the case of supersymmetry. It contains even and uneven elements; the even elements correspond to bosons and the odd elements to fermions. Supersymmetry applies to systems with anticommuting coordinates. Lie superalgebras are a natural generalization of usual Lie algebras to supersymmetry and simply add a Z2-grading.

In detail, a *Lie superalgebra* is a Z2-graded algebra over a field of characteristic 0 – usually **R** or **C** – with a product $[.,.]$, called the (Lie) *superbracket* or *supercommutator*, that has the properties

$$
\begin{aligned}
&[x, y] = -(-1)^{|x||y|}[y, x] \\
&\quad 0 = (-1)^{|z||x|}[x,[y,z]] + (-1)^{|x||y|}[y,[z,x]] + (-1)^{|y||z|}[z,[x,y]]
\end{aligned}
\qquad (833)
$$

where x, y, and z are algebra elements that are 'pure' in the Z2-grading. The expression $|x|$ denotes the degree of the algebra element x and is either 0 or 1. The second condition is sometimes called the super Jacobi identity. Obviously, the even subalgebra of a Lie superalgebra forms a usual Lie algebra; in that case the superbracket becomes the usual Lie bracket.

Like in the case of Lie algebras, also all possible *simple* Lie superalgebras have been classified. They fall into five infinite classes and some special cases, namely

- $A(n, m)$ corresponding to the Lie supergroups $SL(n+1|m+1)$ and their compact 'cousins' $SU(N|M)$,
- $B(n, m)$ corresponding to the Lie supergroups $OSp(2n+1|2m)$,
- $D(n, m)$ corresponding to the Lie supergroups $OSp(2n|2m)$,
- $P(n)$,
- $Q(n)$,
- the exceptional cases $D_\alpha(2,1)$, $G(3)$, $F(4)$,
- and finally the Cartan superalgebrass.

The Virasoro algebra

The *Virasoro algebra* is the infinite algebra of operators L_n satisfying

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}(m^3 - m)\delta_{m,-n} \tag{834}$$

where the number c, which may be zero, is called the *central charge*, and the factor 1/12 is introduced by historical convention. This rather specific algebra is important in physics because it is the algebra of conformal symmetry in *two* dimensions.* Are you able to find a representation in terms of infinite square matrices? Mathematically speaking, the Virasoro algebra is a special case of a Kac–Moody algebra.

Challenge 1516 ny

Kac–Moody algebras

In physics, more general symmetries than Lie groups also appear. This happens in particular when general relativity is taken into account. Due to the symmetries of space-time, the number of generators becomes infinite. The concepts of Lie algebra and of superalgebra have to be extended to take space-time symmetry into account. The related algebras are the Kac-Moody algebras. ('Kac' is pronounced like 'Katz'.) *Kac–Moody algebras* are a particular class of infinite parameter Lie algebras; thus they have an infinite number of generators. In the past they used to be called *associated affine algebras* or *affine Lie algebras*. Sometimes there are also called *Z-graded Lie algebras*. They were introduced independently in 1968 by Victor Kac and Robert Moody. We present some specific examples.

Of basic physical importance are those Kac–Moody algebras associated to a symmetry group $G \otimes C[t]$, where t is some continuous parameter. The generators $M_a^{(n)}$ of the corresponding algebra have two indices, one, a, that numbers the generator of the group G and another, n, that numbers the generators of the group of Laurent polynomials $C[t]$. The generators form a Kac–Moody algebra if they obey the relations:

$$[M_a^{(n)}, M_b^{(m)}] = C_{abc} M_c^{(n+m)} \quad \text{for} \quad n, m = 0, 1, 2, ..., \infty . \tag{835}$$

For example, taking G as SU(2) with its generators $T^n = \sigma^a/2i$, σ_i being the Pauli spin matrices, the objects

$$M_a^{(n)} = T_a \otimes t^n , \quad \text{e.g.} \quad M_3^{(n)} = \frac{1}{2i}\begin{pmatrix} t^n & 0 \\ 0 & -t^n \end{pmatrix} \tag{836}$$

form a Kac–Moody algebra with the structure constants $C_{abc} = \varepsilon_{abc}$ of SU(2).

– CS – more on Kac-Moody algebras to be added – CS –

Page 299

* Note that the conformal symmetry group in *four* dimensions has 15 parameters and thus its Lie algebra is finite (fifteen) dimensional.

Topology – what shapes exist?

Topology is group theory.

The Erlangen program

In a simplified view of topology that is sufficient for physicists, only one type of entities can possess shapes: manifolds. *Manifolds* are generalized examples of pullovers; manifolds are flat, have holes, have boundaries and often can be reversed.

Like manifolds, pullovers are tricky entities. Are you able to find out how to reverse your pullover while your hands are tied to each other? (A friend may help you.) By the way, the same feat is also possible with your trousers, while your feet are tied to each other. Certain professors do this during topology lectures – of course with a carefully selected pair of underpants.

Challenge 1517 n

Topological spaces

Ref. 1176

The study of shapes implies to have a definition of sets made of points. To be able to talk about shapes, these sets must be structured; the structure must define the concept of 'neighbourhood' or 'closeness' between the elements of the set. The search for the most general type of set which allows a useful definition of neighbourhood has lead to the definition of the concept of topological space.

A *topological space* is a finite or infinite set X of elements, called *points* in this case, together with a neighbourhood for each point. A *neighbourhood* N of a point x is a collection of subsets Y_x of X with the properties that

- x is in N,
- If N and M are neighbourhoods, so is $N \cap M$,
- Anything containing a neighbourhood of x is itself a neighbourhood of x.

The choice of the subsets is free. All the subsets Y_x, for all points x, that were chosen in the definition are then called *open sets*. (A neighbourhood and an open set are not necessarily the same; but all open sets are also neighbourhoods.)

One also calls a topological space a 'set with a topology'. In this restricted sense, a *topology* thus is the definition of 'neighbourhood' for every element, or point, of the set. In the wide sense, topology is the name of the part of mathematics which studies topological spaces.

For example, the real numbers together with the open intervals around each point form the *usual* topology of **R**. If one takes *all* subsets containing the point x as neighbourhood of that point, one speaks of the *discrete* topology. If one takes *only* the full set X and the empty set as neighbourhood of each point, one speaks of the *trivial* or indiscrete topology.

The concept of topological space allows to define continuity. A mapping from one topological space X to another topological space Y is *continuous* if the inverse image of an open set in Y is an open set in X. You may test that this definition is not satisfied by a real function that makes a jump. You may also check that the term 'inverse' is necessary in the definition; otherwise a function with a jump would be continuous, as such a function maps open sets into open sets.*

Challenge 1518 e

* The Cauchy-Weierstass definition of continuity says that a real function $f(x)$ is continuous at a point

no figure yet

Figure 377 Examples of orientable and non-orientable manifolds of two dimensions: a disk, a Möbius strip, a sphere and Klein's bottle

We thus need the concept of topological space, or of neighbourhood, if we want to express that nature makes no jumps. We also need the concepts in order to be able to to define limits.

– CS – more on topological spaces to be added – CS –

Many specialized topological spaces have been studied. One type is particularly important. A *Hausdorff space* is a topological space for which for any two points x and y one can define disjunct open sets U and V so that x is in U and y is in V. A Hausdorff space is thus a space where, no matter how close two points are, there always are small enough open sets containing them. This seems obvious; indeed, non-Hausdorff spaces are rather tricky mathematical objects. (Nevertheless, at Planck energy, it seems that vacuum is not a Hausdorff space; on the other hand, at Planck energy, vacuum is not a space either, so that non-Hausdorff spaces play no role in physics.) A special case of Hausdorff space is well-known.

Manifolds

In physics, the most important topological spaces are differential manifolds. Sloppily speaking, a *differential manifold* is a set of points that under the microscope – at small distances – looks like the space $\mathbf{R}^n$. For example, a sphere and a torus are two-dimensional differential manifolds, since they look locally like a plane. However, not all differential manifolds are that simple, as the examples of Figure 377 show.

A differential manifold is called *connected* if any two points can be joined by a path lying in the manifold. (The notion of connectedness and pathwise connectedness coincide for differential manifolds.) We focus on connected manifolds in the following. A manifold is called *simply connected* if every loop lying in the manifold can be contracted to a point. A sphere is simply connected. A connected manifold which is not simply connected is called *multiply connected*. A torus is multiply connected.

Manifolds can be *non-orientable*, such as the well-known Möbius strip makes clear. Non-orientable manifolds have only one surface; they do not distinguish between front and back. If you want to have fun, cut a paper Möbius strip into two along a centre line. You can do this with paper strips with various twists values, and find out the regularities.

Challenge 1520 e

a if (1) f is defined on a open interval containing a, $f(x)$ tends to a limit as x tends to a, and the limit is $f(a)$. In this definition, continuity of f is defined using the continuity of the real numbers. Can you see the connection to the general definition given above?

Challenge 1519 ny

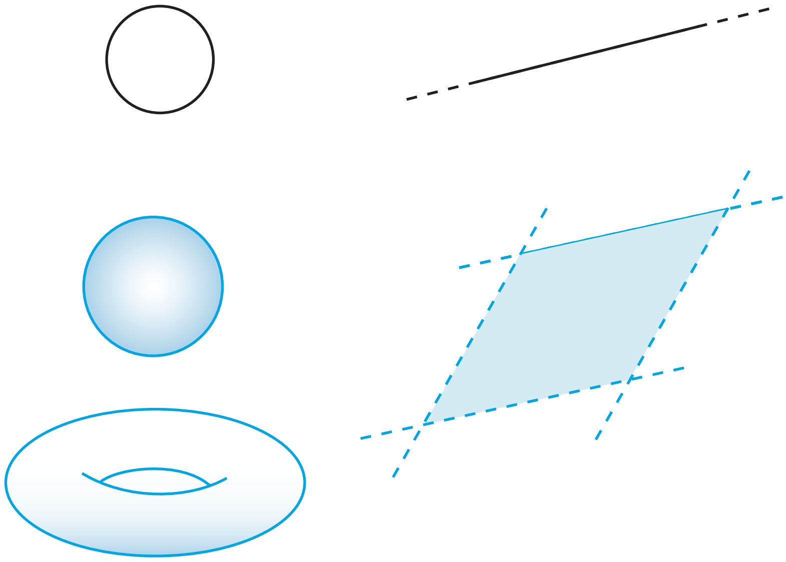


Figure 378 Compact (left) and noncompact (right) manifolds of various dimensions

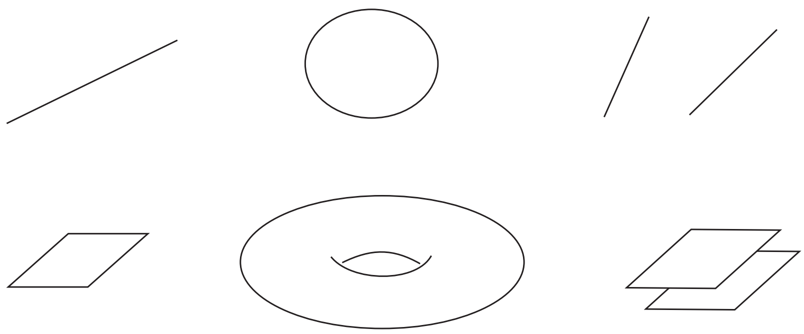


Figure 379 Simply connected (left), multiply connected (centre) and disconnected (right) manifolds of one (above) and two dimensions (below)

In two dimensions, closed manifolds (or surfaces), i.e. surfaces that are compact and without boundary, are always of one of three types:

- The simplest type are spheres with n attached handles; they are called *n-tori* or *surfaces of genus n*. They are orientable surfaces with Euler characteristic $2 - 2n$.
- The *projective planes* with n handles attached are non-orientable surfaces with Euler characteristic $1 - 2n$.
- The Klein bottles with n attached handles are non-orientable surfaces with Euler characteristic $-2n$.

Therefore Euler characteristic and orientability describe a compact surfaces up to homeomorphism (and if surfaces are smooth then up to diffeomorphism).

The two-dimensional *compact* manifolds or surfaces with boundary are found by removing one or more disks from a surface of this list. All compact surfaces can be embedded in $\mathbf{R}^3$ if it is orientable or if it has non-empty boundary.

Figure to be inserted

Figure 380 Examples of homeomorphic pairs of manifolds

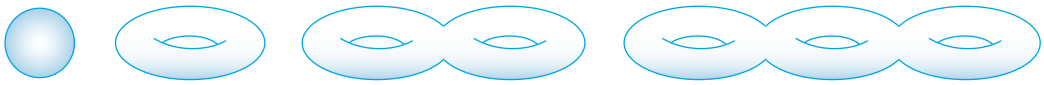


Figure 381 The two-dimensional compact connected orientable manifolds: 0-, 1-, 2-, 3- and n-tori

In physics, the most important manifolds are space-time and Lie groups of observables. We study Lie groups below. Strangely enough, the topology of space-time is not known. For example, it is unclear whether it is simply connected or not. Obviously, the reason is that it is difficult to observe what happens at large distances form the Earth. However, a similar difficulty appears near Planck scales.

If a manifold is imagined to consist of rubber, connectedness and similar global properties are not changed when the manifold is deformed. This fact is formalized by saying that two manifolds are *homeomorphic* (from the Greek for 'same' and 'shape') to each other if between them there is a continuous, one-to-one and onto mapping with a continuous inverse. The concept of homeomorphy is somewhat more general than that of rubber deformation, as can be seen from Figure 380.

Holes, Homotopy and Homology

In physics, only 'kind' manifolds play a role, namely those which are orientable and connected. In addition, for observables, only compact manifolds are found in nature. The main non-trivial characteristic of connected compact orientable manifolds is that they contain 'holes' (see Figure 381). It turns out that a proper description of the holes of manifolds allows to distinguish between all different,i.e. non-homeomorphic types of manifolds.

There are three main tools to describe holes of manifolds and the relations among them: homotopy, homology and cohomology. These tools play an important role in the study of gauge groups, because any gauge group defines a manifold.

– CS – more on topology to be added – CS –

In other words, through homotopy and homology theory, mathematicians can *classify* manifolds. The properties of holes in manifolds thus allow to determine whether two manifolds can be deformed into each other.

Physicists are now extending these results of standard topology. Deformation is a classical idea which assumes continuous space and time, as well as arbitrary small action. In nature however, quantum effects cannot be neglected. It is speculated that quantum ef-

fects can transform a physical manifold into a second one with *different* topology, such as a torus into a sphere. Can you find out how this can be achieved?
Challenge 1521 d

Topology changes of physical manifolds happen via objects that are generalization of manifolds. An *orbifold* is a space that is locally modelled by $\mathbf{R}^n$ modulo a finite group. Examples are the tear-drop or the half-plane. Orbifolds were introduced by Satake Ichiro in 1956 and called orbifolds by William Thurston. Orbifolds are heavily studied in string theory.

Types and classification of groups

Page 176
We introduced groups early on because groups play an important role in many parts of physics, from the description of solids, molecules, atoms, nuclei, elementary particles and forces up to the study of shapes, cycles and patterns in growth processes.

Group theory is also one of the most important branches of modern mathematics, and is still an active area of research. One of the aims of group theory is the *classification* of all possible groups. This program has been achieved only in a few subcases, due to the many types of infinite groups. In general, one distinguishes finite and infinite groups.

All *finite* groups are isomorphic to a subgroup of the symmetric group S_N, with the number N chosen appropriately. Examples of finite groups are the crystalline groups, used to classify crystal structures, or the groups used to classify wallpaper patterns. Also the symmetry groups of platonic and other regular solids are finite groups.

Finite groups are a highly complex family. In a very sloppy way, a general (finite) group can be seen as built from some fundamental bricks, which are groups themselves. These fundamental bricks are called *simple* (finite) groups. One of the high points of twentieth century mathematics is the classification of the simple groups. It was a collaborative effort that took around 30 years, roughly from 1950 to 1980. The complete list of finite simple groups consist of
Ref. 1175

1) the *cyclic groups* Z_p of prime group order;

2) the *alternating groups* A_n with degree n at least five;

3) the *classical linear groups*, $\mathrm{PSL}(n;q)$, $\mathrm{PSU}(n;q)$, $\mathrm{PSp}(2n;q)$ or $\mathrm{P}\Omega^{\varepsilon}(n;q)$;

4) the *exceptional* or *twisted groups* of Lie type ${}^3D_4(q)$, $E_6(q)$, ${}^2E_6(q)$, $E_7(q)$, $E_8(q)$, $F_4(q)$, ${}^2F_4(2^n)0$, $G_2(q)$, ${}^2G_2(3^n)$ or ${}^2B_2(2^n)$;

5) the 26 *sporadic groups*, namely M_{11}, M_{12}, M_{22}, M_{23}, M_{24} (the Mathieu groups), J_1, J_2, J_3, J_4 (the Janko groups), Co_1, Co_2, Co_3 (the Conway groups), HS, Mc, Suz (the Co_1 'babies'), Fi_{22}, Fi_{23}, Fi'_{24} (the Fischer groups), $F_1 = M$ (the Monster), F_2, F_3, F_5, He ($= F_7$) (the Monster 'babies'), Ru, Ly, and ON.

The classification was finished in the 1980s after over 10 000 pages of publications. The amount of work is so vast that special series of books has been started to summarize and explain the proof. The first three families are infinite. The last family, the sporadic group, is the most peculiar; it collects those finite simple groups which do not fit into the infinite families. Some of these sporadic groups might have a role in particle physics, possibly also the largest of them all, the so-called *Monster* group. This issue is still a topic of research. (The monster group has about $8.1 \cdot 10^{53}$ elements; more precisely, its number of elements is 808 017 424 794 512 875 886 459 904 961 710 757 005 754 368 000 000 000 or $2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71$ elements.)

Of all *infinite* groups, only those with some finiteness condition have been studied. Obviously, only such groups are of interest to describe nature. Infinite groups are divided into *discrete* groups and *continuous* groups. Discrete groups are an active area of mathematical research, with connection to number theory and topology. Continuous groups are divided into *finitely generated* and *infinitely generated* groups. Finitely generated groups can be finite dimensional or infinitely dimensional.

The most important class of finitely generated continuous groups are the Lie groups.

Lie groups

In nature, Lagrangians of the fundamental forces are invariant under gauge transformations and under continuous space-time transformations. These symmetry groups are examples of Lie groups, which are special types of infinite continuous groups. They are named after the great Norwegian mathematician Sophus Lie (1849–1899). His name is pronounced like 'Lee'.

A (real) *Lie group* is an infinite symmetry group, i.e. a group with infinitely many elements, which at the same time is also an analytical manifold. Sloppily speaking, this definition means that the elements of the group can at the same time be seen as points on a smooth (hyper-) surface whose shape, i.e. the coordinates of the points of the surface, can be described by an analytic function, i.e. by a function so smooth that it can be expressed as a power series. In addition, the points can be multiplied in a given way, as they are also elements of a group. Furthermore, the coordinates of the product have to be analytical functions of the coordinates of the factors, and the coordinates of the inverse have to be analytic functions of the coordinates of the element itself. In fact, this definition is too strict; it can be proven that a Lie group is any topological group whose topological space is a finite-dimensional, locally Euclidean manifold.

A *complex Lie group* is a group where the manifold is complex and the functions holomorphic instead of analytical.

In short, a Lie group is a nicely behaved manifold in which points can be multiplied (and technicalities). For example, the circle $\mathbf{T} = \{z \in \mathbf{C} : |z| = 1\}$ with usual complex multiplication, is a real Lie group. It is abelian. This group is also called U(1), an abbreviation of 'unitary group of one dimension'. The other one-dimensional Lie groups are the multiplicative group of real numbers and its subgroup, the multiplicative group of positive real numbers.

In physics, only linear Lie groups play a role up to now,i.e. Lie groups which act as linear transformations in some suitable vector space. (The cover of SL(2,**R**) or the complex compact torus are non-linear Lie groups.) The important linear Lie groups for physics are the Lie subgroups of the general linear group GL(N,K), where K is a number field. It is defined as the set of all non-singular, i.e. invertible, N×N real, complex or quaternionic matrices. All Lie groups discussed in the following are of this type.

Every *complex* invertible matrix A can be written in a unique way with help of unitary matrix U and a Hermitean matrix H as

$$A = Ue^H \ . \qquad (837)$$

Challenge 1522 n (H is given as $H = \frac{1}{2} \ln M^\dagger M$, and $U$is given as $U = Me^{-H}$.)

– CS – more on Lie groups to be added – CS –

The Lie groups U(1) and SO(2,**R**) are abelian, all others are non-abelian.

Lie groups are manifolds. As for all manifolds, also for Lie groups one can define the distance between points, the tangent plane (or space) at a point, and the way to integrate and differentiate on the manifold.

Lie groups are manifolds. This means that when every element of the group is seen as a point, Lie groups have the same structure as the objects of Figures 377, 378 and 379. Lie groups can have any number of dimensions. As for any manifold, the details of their global structure contain important information.

Connectedness

Playing around with the group elements, one discovers the following relations. The Lie groups SU(N) are simply connected for all $N = 2, 3...$; they have the topology of a $2N$-dimensional sphere. The Lie Group U(1), having the topology of the 1-dimensional sphere, the circle, is multiply connected.

The Lie groups SO(N) are *not* simply connected for all $N = 2, 3 \ldots$. In general, SO(N,K) is connected, and GL(N,**C**) is connected. The Lie groups SL(N,K) are connected. SL(N,**C**) is simply connected. The Lie groups Sp(N,K) are connected. Sp(2N,**C**) is simply connected. Generally, all semi-simple Lie groups are connected.

The Lie groups O(N,K), SO(N,M,K) and GL(N,**R**) are not connected; they contain two connected components.

As a note, the Lorentz group is not connected; it consists of 4 separate pieces. In addition it is not compact (like the Poincaré group), and neither is any of the 4 pieces. Generally speaking, the non-compactness of space-time symmetries is a consequence of the non-compactness of space-time.

Compactness

A compact Lie group is a group which, when seen as a manifold, is closed and bounded. For a given parametrization of the group elements, the Lie group is *compact* if all parameter ranges are closed and finite intervals; otherwise the group is called *non-compact*. Both compact and non-compact groups play a role in physics. The distinction between the two cases is important, because representations of compact groups can be constructed in the same simple way as for finite groups, whereas for non-compact groups other methods have to be used. As a result, physical observables, which always belong to a representation of a symmetry group, have different properties in the two cases: if the symmetry group is compact, observables have *discrete* spectra, otherwise they do not.

All internal gauge transformations, such as U(1) and SU(N), form compact groups. In fact, field theory *requires* compact Lie groups for gauge transformations. The compact Lie groups are T^N, O(N) and U(N), SO(N) and SU(N). In contrast, SL(N,**R**), GL(N,**R**), GL(N,**C**) are not compact. In fact, all Lie groups except the mentioned five compact classes are not compact.

– CS – more on Lie group representations to be added – CS –

Table 89 Properties of the main real and complex Lie groups

Lie group	Description	Properties	Lie alg.	Description	Dim.
1. Real groups					real
R^n	Euclidean space with addition	abelian, simply connected, not compact	R^n	Lie bracket is zero	n
R^x	nonzero real numbers with multiplication	abelian, not connected, not compact	R	Lie bracket is zero	1
$R > 0$	positive real numbers with multiplication	abelian, simply connected, not compact	R	Lie bracket is zero	1
$S^1 = R/Z$	complex numbers of absolute value 1, with multiplication	abelian, connected, not simply connected, compact	R	Lie bracket is zero	1
H^x	non-zero quaternions with multiplication	simply connected, not compact	H	quaternions, with Lie bracket the commutator	4
S^3	quaternions of absolute value 1, with multiplication, also known as $Sp(1)$; topologically a 3-sphere	simply connected, compact, simple and semi-simple, isomorphic to $SU(2)$ and to $Spin(3)$	$Im(H)$	quaternions with zero real part, with Lie bracket the commutator; isomorphic to real 3-vectors, with Lie bracket the cross product; also isomorphic to $su(2)$ and to $so(3)$	3
$GL(n, R)$	general linear group: invertible n-by-n real matrices	not connected, not compact	$M(n, R)$	n-by-n matrices, with Lie bracket the commutator	n^2
$GL{+}(n, R)$	n-by-n real matrices with positive determinant	simply connected, not compact	$M(n, R)$	n-by-n matrices, with Lie bracket the commutator	n^2

Lie group	Description	Properties	Lie alg.	Description	Dim.
$SL(n, \mathbf{R})$	special linear group: real matrices with determinant 1	simply connected, not compact if $n > 1$	$sl(n, \mathbf{R})$	square matrices with trace 0, with Lie bracket the commutator	$n^2 - 1$
$O(n, \mathbf{R})$	orthogonal group: real orthogonal matrices	not connected, compact	$so(n, \mathbf{R})$	skew-symmetric square real matrices, with Lie bracket the commutator; $so(3, \mathbf{R})$ is isomorphic to $su(2)$ and to $\mathbf{R}^3$ with the cross product	$n(n-1)/2$
$SO(n, \mathbf{R})$	special orthogonal group: real orthogonal matrices with determinant 1	connected, compact, for $n \geqslant 2$: not simply connected, for $n = 3$ and $n \geqslant 5$: simple and semisimple	$so(n, \mathbf{R})$	skew-symmetric square real matrices, with Lie bracket the commutator	$n(n-1)/2$
$Spin(n)$	spinor group	simply connected, compact, for $n = 3$ and $n \geqslant 5$: simple and semisimple	$so(n, \mathbf{R})$	skew-symmetric square real matrices, with Lie bracket the commutator	$n(n-1)/2$
$Sp(2n, \mathbf{R})$	symplectic group: real symplectic matrices	noncompact, simple and semisimple	$sp(2n, \mathbf{R})$	real matrices that satisfy $JA + ATJ = 0$ where J is the standard skew-symmetric matrix	$n(2n+1)$
$Sp(n)$	compact symplectic group: quaternionic $n \times n$ unitary matrices	compact, simply connected, simple and semisimple	$sp(n)$	square quaternionic matrices A satisfying $A = -A^*$, with Lie bracket the commutator	$n(2n+1)$
$U(n)$	unitary group: complex $n \times n$ unitary matrices	isomorphic to S^1 for $n = 1$, not simply connected, compact. (This is not a complex Lie group/algebra)	$u(n)$	square complex matrices A satisfying $A = -A^*$, with Lie bracket the commutator	n^2
$SU(n)$	special unitary group: complex $n \times n$ unitary matrices with determinant 1	simply connected, compact, for $n \geqslant 2$: simple and semisimple. (This is not a complex Lie group/algebra)	$su(n)$	square complex matrices A with trace 0 satisfying $A = -A^*$, with Lie bracket the commutator	$n^2 - 1$
2. Complex groups					complex

Lie group	Description	Properties	Lie alg.	Description	Dim.
$\mathbb{R}^n$	group operation is addition	abelian, simply connected, not compact	$\mathbb{R}^n$	Lie bracket is zero	n
$\mathbb{R}^x$	nonzero complex numbers with multiplication	abelian, not simply connected, not compact	$\mathbb{R}$	Lie bracket is zero	1
$GL(n,\mathbb{R})$	general linear group: invertible n-by-n complex matrices	simply connected, not compact, for $n = 1$: isomorphic to $\mathbb{R}^x$	$M(n,\mathbb{R})$	n-by-n matrices, with Lie bracket the commutator	n^2
$SL(n,\mathbb{R})$	special linear group: complex matrices with determinant 1	simple, semisimple, simply connected, for $n \geqslant 2$: not compact	$sl(n,\mathbb{R})$	square matrices with trace 0, with Lie bracket the commutator	$n^2 - 1$
$O(n,\mathbb{R})$	orthogonal group: complex orthogonal matrices	not connected, for $n \geqslant 2$: not compact	$so(n,\mathbb{R})$	skew-symmetric square complex matrices, with Lie bracket the commutator	$n(n-1)/2$
$SO(n,\mathbb{R})$	special orthogonal group: complex orthogonal matrices with determinant 1	for $n \geqslant 2$: not compact, not simply connected, for $n = 3$ and $n \geqslant 5$: simple and semisimple	$so(n,\mathbb{R})$	skew-symmetric square complex matrices, with Lie bracket the commutator	$n(n-1)/2$
$Sp(2n,\mathbb{R})$	symplectic group: complex symplectic matrices	noncompact, simple and semisimple	$sp(2n,\mathbb{R})$	complex matrices that satisfy $JA + ATJ = 0$ where J is the standard skew-symmetric matrix	$n(2n+1)$

This table is found in the Wikipedia, on the http://en.wikipedia.org/wiki/Lie_group website.

Besides being manifolds, Lie groups obviously are also groups. It turns out that most of their group properties can be studied by investigating the behaviour of the elements which are very close to the identity, when the group elements are seen as points on the manifold.

Investigating more closely, one finds that for a compact and connected Lie group G, every element g of G has the form $g = \exp(A)$ for some $A \in L$. The elements A form an algebra L, this algebra is called the corresponding *Lie algebra*. For any linear Lie group, every element of the connected subgroup can be expressed as finite product of exponentials of the real Lie algebra L. In short, Lie algebras are the local properties of Lie groups. That is the reason that they were explored above.

Page 1112

Mathematical curiosities and fun challenges

Mathematics is a topic for a passion all by itself.

▪ Mathematics provides many counter-intuitive results. Reading a book on the topic, such as Bernard R. Gelbaum & John M.H. Olmsted, *Theorems and Counter-examples in Mathematics*, Springer, 1993, can help you sharpen your mind.

▪ The distinction between one, two and three dimensions is weak in mathematics. This is best shown in the text Hans Sagan, *Space Filling Curves*, Springer Verlag, 1994.

▪ There are at least seven ways to win one million dollar with mathematical research. The Clay Mathematics Institute at http://www.claymath.org offers them for advances in seven topics:

- proving the Birch and Swinnerton–Dyer conjecture about algebraic equations;
- proving the Poincaré conjecture about topological manifolds;
- solving the Navier–Stokes equations for fluids;
- finding criteria distinguishing P and NP numerical problems;
- proving the Riemann hypothesis stating that the non-trivial zeros of the zeta function lie on a line;
- proving the Hodge conjectures;
- proving the connection between Yang–Mills theories and a mass gap in quantum field theory.

On each of these topics, substantial progress can buy you a house.

> No man but a blockhead ever wrote except for money.
>
> Samuel Johnson

Bibliography

1167 A general basis can be the *Encyclopedia of Mathematics*, in 10 volumes, Kluwer Academic Publishers, Dordrecht, 1988–1993. It explains carefully all concepts used in mathematics. Spending an hour with it looking up related keywords is an efficient way to get an introduction into any part of mathematics, especially into the vocabulary and the main connections. Cited on page 1098.

1168 S.L. Altman, Rotations, Quaternions and Double Groups, Clarendon Press, Oxford, 1986, and also S.L. Altman, Hamilton, Rodriguez and the quaternion scandal, *Mathematical Magazine* pp. 291–308, 1988. Cited on page 1103.

1169 A. Waser, Quaternions in Electrodynamics, 2001. The text can be downloaded at the website http://www.aw-verlag.ch/Documents/QuaternionsInElectrodynamicsEN02.pdf. His article http://www.aw-verlag.ch/Documents/NotationofMaxwellFieldEquations.PDF is a clear explanation of the history of the notation. Cited on pages 1101 and 1104.

1170 See the fine book by Louis H. Kauffman, *Knots and Physics*, World Scientific, second edition, 1994, which gives a clear and visual introduction to the mathematics of knots and its main applications to physics. Cited on page 1103.

1171 A good introduction to nonstandard numbers, quaternions, octonions, p-adic numbers, surreal numbers and more is H.-D. Ebbinghaus, H. Hermes, F. Hirzebruch, M. Koecher, K. Mainzer, J. Neukirch, A. Prestel & R. Remmert, *Zahlen*, Springer Verlag, 1993, also available in English as *Numbers*, Springer Verlag, 1990. Cited on pages 1106 and 1107.

1172 Gaussian integers are explored by G.H. Hardy & E.M. Wright, *An Introduction to the Theory of Numbers*, 5th ed. Clarendon Press, Oxford, 1979, in the section 12.2 The Rational Integers, the Gaussian Integers, and the Integers of, pp. 178–180,, and 12.6 Properties of the Gaussian Integers, pp. 182–183. For their challenges, look at http://www.mathpuzzle.com/Gaussians.html. Cited on page 1107.

1173 About transfinite numbers, see the delightful paperback by Rudy Rucker, *Infinity and the Mind – the Science and Philosophy of the Infinite*, Bantam, Toronto, 1983. Cited on page 1107.

1174 M. Flato, P. Sally & G. Zuckerman, (editors) Applications of Group Theory in Physics and Mathematical Physics, Lectures in applied mathematics, volume 21, American Mathematical Society 1985. This interesting book has been written before the superstring revolution, so that the latter topic is missing in the otherwise excellent presentation. Cited on page 1114.

1175 An introduction on the classification theorem is R. Solomon, On finite simple groups and their classification, *Notices of the AMS* 42, pp. 231–239, 1995, also to be found on the web as http://www.ams.org/notices/199502/solomon.ps Cited on page 1120.

1176 For an introduction to topology, see for example Mikio Nakahara, *Geometry, Topology and Physics*, IOP Publishing, Bristol 1990. Cited on page 1116.

Appendix E

Information Sources on Motion

No place affords a more striking conviction of the vanity of human hopes than a public library.
Samuel Johnson (1709–1784)

In a consumer society there are inevitably two kinds of slaves: the prisoners of addiction and the prisoners of envy.
Ivan Illich (b. 1926 Vienna, d. 2002 Bremen)

In the text, outstanding books introducing neighbouring domains are presented in footnotes. The reference list at the end of each chapter collects general material satisfying further curiosity about what is encountered in this mountain ascent. All citations can also be found by looking up the author's name in the index. To find additional information, either libraries or the internet can help.

In a library, review articles of recent research appear in journals such as *Reviews of Modern Physics*, *Reports on Progress in Physics*, *Contemporary Physics* and *Advances in Physics*. Pedagogical introductions are best found in the *American Journal of Physics*, the *European Journal of Physics* and in *Physik in unserer Zeit*. A special case are the *Living Reviews in Relativity* to be found at http://www.livingreviews.org.

Actual overviews on research trends can be found irregularly in magazines such as *Physics World*, *Physics Today*, *Europhysics Journal*, *Physik Journal* and *Nederlands Tijdschrift voor Natuurkunde*. For all sciences together, the best sources are the magazines *Nature*, *New Scientist*, *Naturwissenschaften*, *La Recherche* and the cheap but excellent *Science News*.

Research papers appear mainly in *Physics Letters B*, *Nuclear Physics B*, *Physical Review D*, *Physical Review Letters*, *Classical and Quantum Gravity*, *General Relativity and Gravitation*, *International Journal of Modern Physics* and in *Modern Physics Letters*. The newest results and speculative ideas are found in conference proceedings, such as the *Nuclear Physics B Supplements*. Articles on the topic can also appear in *Fortschritte der Physik*, *Zeitschrift für Physik C*, *La Rivista del Nuovo Cimento*, *Europhysics Letters*, *Communications in Mathematical Physics*, *Journal of Mathematical Physics*, *Foundations of Physics*, *International Journal of Theoretical Physics* and *Journal of Physics G*. There is also the purely electronic *New Journal of Physics* that can be found at the http://www.njp.org website.

Papers on the description of motion without time and space which appear *after* this text is published can be found via the *Scientific Citation Index*. It is published in printed form or as compact disk and allows, given a paper, e.g. one from the references at the end of each chapter, to search for all subsequent publications which cite it. Then, using the

Table 90 The structure of the Arxiv preprint archive for physics and related topics at http://www.arxiv.org

Topic	Abbreviation
general relativity and quantum cosmology	gr-qc
astrophysics	astro-ph
experimental nuclear physics	nucl-ex
theoretical nuclear physics	nucl-th
theoretical high energy physics	hep-th
computational high energy physics	hep-lat
phenomenological high energy physics	hep-ph
experimental high energy physics	hep-ex
quantum physics	quant-ph
general physics	physics
condensed matter physics	cond-mat
nonlinear sciences	nlin
mathematical physics	math-ph
mathematics	math
computer science	CoRR
quantitative biology	q-bio

To receive preprints per email, send a mail to the addresses in the form mailto:gr-qc@arxiv.org (or the corresponding abbreviation) and use a subject line consisting simply of the word 'help', without the quotes.

bimonthly *Physics Abstracts*, which also exists both in paper and in electronic form, you can look up the abstract of the paper and check whether it is of interest.

But by far the simplest and most efficient way to keep in touch with ongoing research on motion is with the help of the *internet*, the international computer network. To anybody with a personal computer connected to a telephone, most theoretical physics papers are available free of charge, as *preprints*, i.e. before official publication and check by referees. This famous service is available at the http://www.arxiv.org website. A service for finding subsequent preprints that cite a given one is also available.

In the last decade of the twentieth century, the internet expanded into a mix of library, media store, discussion platform, order desk, brochure collection and time waster. With a personal computer, a modem and free *browser* software one can look for information in millions of pages of documents. The various parts of the documents are located in various computers around the world, but the user does not need to be aware of this.*

* Decades ago, the provoking book by Ivan Illich, *Deschooling Society*, Harper & Row, 1971, listed four basic ingredients for any educational system:

1. access to *resources* for learning, e.g. books, equipment, games, etc. at an affordable price, for everybody, at any time in their life;

2. for all who want to learn, access to *peers* in the same learning situation, for discussion, comparison,

To start using the web, ask a friend who knows, or send an electronic mail message consisting of the line 'HELP' to listserv@info.cern.ch, the server at the European Organization for Particle Research, where the web was invented.* Searching the web for authors, organizations, books, publications, companies or simple keywords using search engines can be a rewarding or a time-wasting experience. A selection of interesting servers are given below.

Table 91 Some interesting world wide web servers

Topic	Website address or 'URL'
Search engines and more	
Good information search engines	http://www.altavista.com
	http://www.metager.de
	http://www.google.com
Search old usenet articles	http://groups.google.com
Information about the net	http://akebono.stanford.edu/yahoo/
	http://cuiwww.unige.ch/w3catalog
Frequently asked questions on various topics, also on physics	http://www.faqs.org
Physics and science	
Electronic preprints	http://www.arxiv.org and others – see above
	http://www.slac.stanford.edu/spires
High energy physics	http://mentor.lanl.gov/Welcome.html or http://info.cern.ch/ hypertext/DataSources/bySubject/Physics/ HEP.html
Particle data	http://pdg.web.cern.ch/pdg
Physics news, weekly	http://www.aip.org/physnews/update

cooperation, competition;

3. access to *elders*, e.g. teachers, for their care and criticism towards those who are learning;

4. exchanges between student and *performers* in the field of interest, so that the latter can be models to the former. For example, there should be the possibility to listen to professional musicians, reading the works of specialist writers, as well as giving performers the possibility to share, to advertise and to perform their skills.

Illich develops the idea that if such a system were informal – he then calls it a 'learning web' or 'opportunity web' – it would be superior to formal, state financed institutions, such as existing schools, for the development of mature human beings. The discussion is deepened in his following works, *Deschooling Our Lives*, Penguin, 1976, and Tools for Conviviality, Penguin, 1973. Today, any networked computer offers one or more of the following: *e-mail* (electronic mail), *ftp* (file transfer to and from another computer), access to *usenet* (the discussion groups on specific topics, such as particle physics) and the powerful *world-wide web*. (Simply speaking, each of the latter implies and includes the ones before.) In a rather unexpected way, all these facilities of the internet have transformed it into the backbone of the opportunity web mentioned by Illich. However, like in real schools, it strongly depends on the user's discipline whether the world wide web actually does provide a learning web.

* To use ftp via electronic mail, send a message to mailto:archie@archie.mcgill.ca with 'help' as mail text. To get web pages via e-mail, send an e-mail message to w3mail@gmd.de consisting of the word 'help', or, for general instructions, send it to mail-server@rtfm.mit.edu with the mail text 'send usenet/news.answers/internet-services/access-via-email'.

Topic	Website address or 'URL'
Physics problems by Yakov Kantor	http://star.tau.ac.il/QUIZ/
Physics problems by Henry Greenside	http://www.phy.duke.edu/~hsg/physics-challenges/challenges.html
Physics question of the week	http://www.physics.umd.edu/lecdem/outreach/QOTW/active
Miniproblem	http://www.nyteknik.se/miniproblemet
Official SI unit site	http://www.bipm.fr
Unit conversion	http://www.chemie.fu-berlin.de/chemistry/general/units-en.html
Ask the expert	http://www.sciam.com/askexpert_directory.cfm
Article summaries in 25 science magazines	http://www.mag.browse.com/science.html
Abstracts of papers in physics journals	http://www.osti.gov
Science News	http://www.sciencenews.org
Nobel Prize winners	http://www.nobel.se/physics/laureates
Pictures of physicists	http://www.if.ufrj.br/famous/physlist.html
Information on physicists	http://144.26.13.41/phyhist
Gravitation news	http://vishnu.nirvana.phys.psu.edu/mog.html
Living reviews in relativity	http://www.livingreviews.org
Information on relativity	http://math.ucr.edu/home/baez/relativity.html
Relativistic imaging and movies	http://www.tat.physik.uni-tuebingen.de/~weiskopf
Physics problems	http://star.tau.ac.il/QUIZ
Physics organizations	http://www.cern.ch/
	http://info.cern.ch/
	http://aps.org
	http://www.hep.net/documents/newsletters/pnu/pnu.html
	http://www.aip.org
	http://www.nikhef.nl/www/pub/eps/eps.html
	http://www.het.brown.edu/physics/review/index.html
Physics textbooks on the web	http://www.motionmountain.net
	http://www.plasma.uu.se/CED/Book
	http://www.biophysics.org/education/resources.htm
	http://www.lightandmatter.com
Three beautiful French sets of notes on classical mechanics and particle theory	http://www.phy.ulaval.ca/enote.html
Th excellent *Radical Freshman Physics* by David Raymond	http://www.physics.nmt.edu/~raymond/teaching.html
Physics lecture scripts in German and English	http://kbibmp5.ub.uni-kl.de/Linksammlung/Physik/liste.html

Topic	Website address or 'URL'
'World' lecture hall	http://www.utexas.edu/world/lecture
Engineering data and formulae	http://www.efunda.com/
Math forum internet resource collection	http://mathforum.org/library/
Biographies of mathematicians	http://mathforum.org/library/
Purdue math problem of the week	http://www.math.purdue.edu/academics/pow/
Macalester college math problem of the week	http://mathforum.org/wagon/
Math formulae	http://dlmf.nist.gov
Functions	http://functions.wolfram.com
Symbolic integration	http://www.integrals.com
	http://http.cs.berkeley.edu/~fateman/htest.html
Weisstein's World of Mathematics	http://mathworld.wolfram.com
Libraries	http://www.konbib.nl
	http://portico.bl.uk
	http://portico.bl.uk/gabriel/en/services.html
	http://www.niss.ac.uk/reference//opacsalpha.html
	http://www.bnf.fr
	http://www.laum.uni-hannover.de/iln/bibliotheken/kataloge.html
	http://www.loc/gov
	http://lcweb.loc.gov
Publishers	http://www.ioppublishing.com/
	http://www.aip.org
	http://www.amherts.edu/~ajp
	http://www.elsevier.nl/
	http://www.nature.com/
Computers	
File conversion	http://tom.cs.cmu.edu/intro.html
Download software and files	http://www.filez.com
Curiosities	
Minerals	http://webmineral.com
	http://www.mindat.org/com
ESA	http://sci.esa.int
NASA	http://oel-www.jpl.nasa.gov/basics/bsf.html
	http://gsfc.nasa.gov
Hubble space telescope	http://hubble.nasa.gov
The cosmic mirror	http://www.astro.uni-bonn.de/~dfischer/mirror

Topic	Website address or 'URL'
Solar system simulator	http://space.jpl.nasa.gov
Observable satellites	http://liftoff.msfc.nasa.gov/RealTime/JPass/20/
Astronomy picture of the day	http://antwrp.gsfc.nasa.gov/apod/astropix.html
The Earth from space	http://www.visibleearth.nasa.gov
Current solar data	http://www.n3kl.org/sun
Optical illusions	http://www.sandlotscience.com
Petit's science comics	http://www.jp-petit.com/science/index.html
Physical toys	http://www.e20.physik.tu-muenchen.de/~cucke/toylink.htm
Physics humour	http://www.escape.ca/~dcc/phys/humor.htm
Literature on magic	http://www.faqs.org/faqs/magic-faq/part2/
Algebraic surfaces	http://www.mathematik.uni-kl.de/~{}hunt/drawings.html
Making paper aeroplanes	http://pchelp.inc.net/paper_ac.htm
Small flying helicopters	http://pixelito.reference.be
	http://www.ivic.qc.ca/~aleexpert/aluniversite/ klinevogelmann.html
Ten thousand year clock	http://www.longnow.org
Gesellschaft Deutscher Naturforscher und Ärzte	http://www.gdnae.de/
Pseudoscience	http://suhep.phy.syr.edu/courses/modules/PSEUDO/pseudo_main.html
Crackpots, English language	http://www.crank.net
Mathematical quotations	http://math.furman.edu/~mquot.html
The World Question Center	http://www.edge.org
Hoaxes	http://www.museumofhoaxes.com

Do you want to study physics without actually going to university? Nowadays it is possible to do so via e-mail and internet, in German, at the University of Kaiserslautern.* In the near future, a nationwide project in Britain should allow the same for English speaking students. As introduction, use the last update of this physics text!

> Das Internet ist die offenste Form der geschlossenen Anstalt.**
>
> Matthias Deutschmann

> Si tacuisses, philosophus mansisses.***
>
> After Boethius.

* See the http://www.fernstudium-physik.de website.
** 'The internet is the most open form of a closed institution.'
*** 'If you had kept quiet, you would have remained philosopher.' After the story Boethius tells in *De consolatione philosophiae*, 2,7, 67 ff.

Appendix F

Challenge Hints & Solutions

> Never make a calculation before you know the answer.
>
> John Wheeler's motto

John Wheeler wanted people to estimate, to try and to guess; but not saying it out loud. A correct guess reinforces the physics instinct, whereas a wrong one leads to the pleasure of surprise. This text contains 1522 challenges. Let me know the challenge for which you want a hint or a solution to be added next.

Challenge 2, page 23: These topics are all addressed later in the text.

Challenge 3, page 30: There are many ways to distinguish real motion from an illusion of motion: for example, only real motion can be used to set something else into motion. In addition, the motion illusion of the figures shows an important failure; nothing moves if the head and the paper remain fixed with respect to each other. In other words, the illusion only *amplifies* existing motion, it does not *create* motion from nowhere.

Challenge 4, page 30: Without detailed and precise experiments, both sides can find examples to prove their point. Creation is supported by the appearance of mould or bacteria in a glass of water; creation is supported by its opposite, namely traceless disappearance, such as the disappearance of motion. Conservation is supported by all those investigations which look into assumed cases of appearance or disappearance.

Challenge 6, page 32: Political parties, sects, helping organizations and therapists of all kinds are typical for this behaviour.

Challenge 7, page 35: The issue is not yet completely settled for the motion of empty space, such as in the case of gravitational waves. For sure, empty space is not made of small particles of finite size, as this would contradict the transversality of gravity waves.

Challenge 8, page 37: It is: objects are defined as what moves with respect to the background, and the background is defined as what stays when objects change. We shall return to this important issue several times in our adventure. It will require a lot of patience to solve it, though.

Challenge 10, page 39: See page 760.

Challenge 11, page 39: A ghost can be a moving image; it cannot be a moving object, as objects cannot interpenetrate. See page 732.

Challenge 12, page 39: Hint: yes, there is such a point.

Challenge 13, page 39: Can one show at all that something has stopped moving?

Challenge 14, page 39: How would you measure this?

Challenge 15, page 39: The number of reliable digits of a measurement result is a simple quantification of precision.

Challenge 16, page 39: No; memory is needed for observation and measurements.

Challenge 17, page 39: Note that you never have observed zero speed.

Challenge 18, page 40: $(2^{64} - 1) = 18\,446\,744\,073\,700\,551\,615$ grains of rice, given a world harvest of 500 million tons, are about 4000 years of rice harvests.

Challenge 19, page 40: Some books state that the flame leans inwards. But experiments are not easy, and sometimes the flame leans outwards. Just try it. Can you explain your observations?

Challenge 20, page 40: Accelerometers are the simplest devices. They exist as piezo devices that produce a signal whenever the box is accelerated and can cost as little as one Euro. Another accelerometer that might have a future is an interference accelerometer that makes use of the motion of an interference grating; this device might be integrated in silicon. Other, more precise accelerometers use gyroscopes or laser beams running in circles.

Velocimetersand position detectors can also detect motion; they need a wheel or at least an optical way to look out of the box. Tachographs in cars are examples of velocimeters, computer mice are examples of position detectors.

Challenge 21, page 40: It rolls towards the centre of the table, as the centre is somewhat lower than the border, shoots over, and then performs an oscillation around that centre. The period is 84 min, as shown in challenge 290.

Challenge 22, page 40: Accelerations can be felt. Many devices measure accelerations and then deduce the position. They are used in aeroplanes when flying over the atlantic.

Challenge 23, page 40: The necessary rope length is nh, where n is the number of wheels/pulleys.

Challenge 24, page 40: The block moves twice as fast as the cylinders, independently of their radius.

Challenge 25, page 40: This methods is known to work with other fears as well.

Challenge 26, page 40: Three couples require 11 passages. Two couples require 5. For four or more couples there is no solution. What is the solution if there are n couples and $n - 1$ places on the boat?

Challenge 27, page 40: In everyday life, this is correct; what happens when quantum effects are taken into account?

Challenge 28, page 42: There is only one way: compare it with the speed of light.

Challenge 29, page 43: The average distance change of two neighbouring atoms in a piece of quartz over the last million years. Do you know something still slower?

Challenge 30, page 44: Equivalently: do points in space exist? The third part studies this issue in detail; see page 923.

Challenge 31, page 44: All electricity sources must use the same phase when they feed electric power into the net. Clocks of computers on the internet must be synchronized.

Challenge 32, page 44: Note that the shift increases quadratically with time, not linearly.

Challenge 33, page 45: Natural time is measured with natural motion. Natural motion is the motion of light. Natural time is this defined with the motion of light.

Challenge 34, page 46: Galileo measured time with a scale. His watch was a water tube that he kept closed with his thumb, pointing into a bucket. To start the stopwatch, he removed his thumb, to stop it, he put it back on. The volume of water in the bucket then gave him a measure of the time interval. This is told in his famous book Galileo Galilei, *Discorsi e dimostrazioni matematiche intorno a due nuove scienze attenenti alla mecanica e i movimenti locali*, usually simply called the 'Discorsi', which he published in 1638 with Louis Elsevier in Leiden, in the Netherlands.

Challenge 35, page 47: There is no way to define a local time at the poles that is consistent with neighbouring points.

Challenge 37, page 48: The forest is full of light and thus of light rays.

Challenge 38, page 49: One pair of muscles moves the lens along the third axis by deforming the eye from prolate to spherical to oblate.

Challenge 39, page 49: This you can solve trying to think in four dimensions. Try to imagine how to switch the sequence when two pieces cross.

Challenge 40, page 50: Light.

Challenge 43, page 52: It is easier to work with the unit torus. Take the unit interval $[0, 1]$ and equate the end points. Define a set B in which the elements are a given real number b from the interval plus all those numbers who differ from that real by a rational number. The unit circle can be thought as the union of all the sets B. (In fact, every set B is a shifted copy of the rational numbers $\mathbf{Q}$.) Now build a set A by taking one element from each set B. Then build the set family consisting of the set A and its copies A_q shifted by a rational q. The union of all these sets is the unit torus. The set family is countably infinite. Then divide it into *two* countably infinite set families. It is easy to see that each of the two families can be renumbered and its elements shifted in such a way that each of the two families forms a unit torus.

Ref. 32 Mathematicians say that there is no countably infinitely additive measure of $\mathbf{R}^n$ or that sets such as A are non-measurable. As a result of their existence, the 'multiplication' of lengths is possible. Later on we shall explore whether bread or *gold* can be multiplied in this way.

Challenge 44, page 52: This puzzle is left to solve to the reader.

Challenge 45, page 53: An example is the region between the x-axis and the function which assigns 1 to every transcendental and 0 to every non-transcendental number.

Challenge 46, page 53: We use the definition of the function of the text. The dihedral angle of a regular tetrahedron is an irrational multiple of π, so the tetrahedron has a non-vanishing Dehn invariant. The cube has a dihedral angle of $\pi/2$, so the Dehn invariant of the cube is 0. Therefore, the cube is not equidecomposable with the regular tetrahedron.

Challenge 47, page 54: If you think you can show that empty space is continuous, you are wrong. Check your arguments. If you think you can prove the opposite, you *might* be right – but only if you already know what is explained in the third part of the text.

Challenge 48, page 54: Obviously, we use light to check that the plumb line is straight, so the two definitions must be the same. This is the case because the field lines of gravity are also possible paths for the motion of light. However, this is not always the case; can you spot the exceptions?

Another way to check straightness is along the surface of calm water.

Challenge 49, page 55: The hollow Earth theory is correct if the distance formula is used consistently. In particular, one has to make the assumption that objects get smaller as they approach the centre of the hollow sphere. Good explanations of all events are found on http://www.geocities.com/inversedearth/ Quite some material can be found on the internet, also under the names of celestrocentric system, inner world theory or concave Earth theory. There is no way to prefer one description over the other, except possibly for reasons of simplicity or intellectual laziness.

Challenge 53, page 56: The required gap is

$$d = \sqrt{(L-b)^2 - w^2 + 2w\sqrt{R^2 - (L-b)^2}} - L + b \ ,$$

as deduced from Figure 382.

Challenge 54, page 56: A smallest gap does not exist: any value will do! Can you show this?

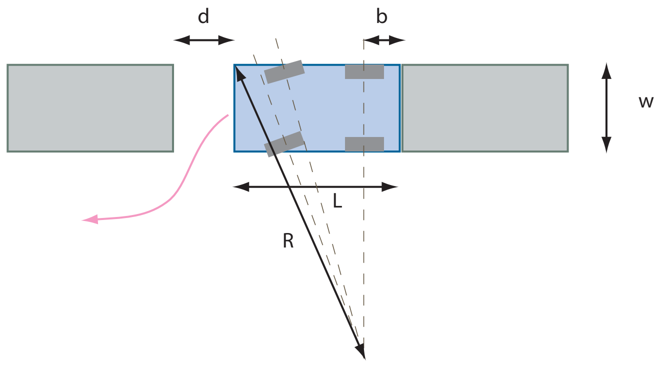


Figure 382 Leaving a parking space – the outer turning radius

Challenge 55, page 56: The first solution sent in will go here.

Challenge 56, page 57: 22 times. 2 times.

Challenge 57, page 57: For two hands, the answer is 143 times.

Challenge 58, page 57: The Earth rotates with 15 minutes per minute.

Challenge 59, page 57: You might be astonished, but no reliable data exist on this question. The highest speed of a throw measured so far seems to be a 45 m/s cricket bowl. By the way, much more data are available for speeds achieved with the help of rackets. The *c.* 70 m/s of fast badminton smashes seem to be a good candidate for record racket speed; similar speeds are achieved by golf balls.

Challenge 60, page 57: Yes, it can. In fact, many can do so at the same time.

Challenge 61, page 57: 1.8 km/s.

Challenge 62, page 57: Nothing, neither a proof nor a disproof.

Challenge 63, page 57: The different usage reflects the idea that we are able to determine our position by ourselves, but not the time in which we are. The section on determinism will show how wrong this distinction is. Page 141

Challenge 64, page 57: Yes, there is. However, this is not obvious, as it implies that space and time are not continuous, in contrast to what we learn in primary school. The answer will be found in the third part of this text.

Challenge 65, page 57: For a curve, use, at each point, the curvature radius of the circle approximating the curve in that point; for a surface, define two directions in each point and use two such circles along these directions.

Challenge 66, page 57: It moves about 1 cm in 50 ms.

Challenge 68, page 58: The final shape is a full cube without any hole.

Challenge 69, page 58: See page 253.

Challenge 70, page 58: A hint for the solution is given by Figure 383.

Challenge 71, page 58: Because they are or were fluid.

Challenge 72, page 58: The shape is shown in Figure 384; it has eleven lobes.

Challenge 73, page 58: The cone angle φ is related to the solid angle Ω through $\Omega = 2\pi(1 - \cos\varphi/2)$.

Challenge 76, page 59: Hint: draw all objects involved.

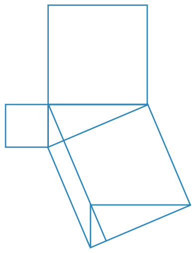


Figure 383 A simple drawing proving Pythagoras' theorem

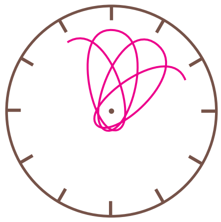


Figure 384 The trajectory of the middle point between the two ends of the hands of a clock

Challenge 77, page 59: Hint: there are an infinite number of shapes.

Challenge 78, page 59: The curve is obviously called a catenary, from Latin 'catena' for chain. The formula for a catenary is $y = a \cosh(x/a)$. If you approximate the chain by short straight segments, you can make wooden blocks that can form an arch without any need for glue. The St. Louis arch is in shape of a catenary. A suspension bridge has the shape of a catenary before it is loaded, i.e. before the track is attached to it. When the bridge is finished, the shape is in between a catenary and a parabola.

Challenge 79, page 60: A limit does not exist in classical physics; however, there is one in nature which appears as soon as quantum effects are taken into account.

Challenge 80, page 60: The inverse radii, or curvatures, obey $a^2 + b^2 + c^2 + d^2 = (1/2)(a + b + c + d)^2$. This formula was discovered by René Descartes. If one continues putting circles in the remaining spaces, one gets so-called circle packings, a pretty domain of recreational mathematics. They have many strange properties, such as intriguing relations between the coordinates of the circle centres and their curvatures.

Challenge 82, page 60: Draw a logarithmic scale, i.e., put every number at a distance corresponding to its natural logarithm.

Challenge 83, page 60: Two more.

Challenge 84, page 60: The Sun is exactly behind the back of the observer; it is setting, and the rays are coming from behind and reach deep into the sky in the direction opposite to that of the Sun. A slightly different situation appears when you have a lighthouse in your back. Can you draw it?

Challenge 86, page 61: From $x = gt^2/2$ you get the following rule: square the number of seconds, multiply by five and you get the depth in metres.

Challenge 87, page 62: Just experiment.

Challenge 88, page 62: The Academicians suspended one cannon ball with a thin wire just in front of the mouth of the cannon. When the shot was released, the second, flying cannon ball flew through the wire, thus ensuring that both balls started at the same time. An observer from far away then tried to determine whether both balls touched the Earth at the same time. The experiment is not easy, as small errors in the angle and air resistance confuse the results.

Challenge 89, page 62: A parabola has a so-called focus or focal point. All light emitted from that point and reflected exits in the same direction: all light ray are emitted in parallel. The name 'focus' – Latin for fireplace – expresses that it is the hottest spot when a parabolic mirror is illuminated. Where is the focus of the parabola $y = 2x$? (Ellipses have two foci, with a slightly different definition. Can you find it?)

Challenge 90, page 62: Neglecting air resistance and approximating the angle by 45°, we get $v = \sqrt{dg}$, or about 3.8 m/s. This speed is created by a stead pressure build-up, using blood pressure, which is suddenly released with a mechanical system at the end of the digestive canal. The cited reference tells more about the details.

Challenge 91, page 63: On horizontal ground, for a speed v and an angle from the horizontal α, neglecting air resistance and the height of the thrower, the distance d is $d = v^2 \sin 2\alpha/g$.

Challenge 92, page 63: Walk or run in the rain, measure your own speed v and the angle from the vertical α with which the rain appears to fall. Then the speed of the rain is $v_{\text{rain}} = v/\tan\alpha$.

Challenge 93, page 63: Check your calculation with the information that the 1998 world record is juggling with nine balls.

Challenge 94, page 63: The long jump record could surely be increased by getting rid of the sand stripe and by measuring the true jumping distance with a photographic camera; that would allow jumpers to run more closely to their top speed. The record could also be increased by a small inclined step or by a spring-suspended board at the take-off location, to increase the take-off angle.

Challenge 96, page 63: It seems not too much. But the lead in them can poison the environment.

Challenge 95, page 63: It is said so, as rain drops would then be ice spheres and fall with high speed.

Challenge 97, page 63: Stones do not fall parabolas when studied in detail, i.e. when the change of g with height is taken into account. Their precise path is an ellipse. The shape appears for long throws, such as throws around the part of the Earth, or for orbiting objects. In short, stones follow parabolas only if the Earth is assumed to be flat. If its curvature is taken into account, they follow ellipses.

Challenge 100, page 66: The set of all rotations around a point in a plane is a vector space. What about the set of all rotations around *all* points in a plane? And what about the three-dimensional cases?

Challenge 101, page 66: The scalar product between two vectors **a** and **b** is given by $ab \cos \sphericalangle(\mathbf{a}, \mathbf{b})$. How does this differ form the vector product?

Challenge 103, page 68: One can argue that any source of light must have finite size.

Challenge 105, page 68: What the unaided human eye perceives as a tiny black point is usually about 50 μm in diameter.

Challenge 106, page 68: See page 526.

Challenge 107, page 68: One has to check carefully whether the conceptual steps that lead us to extract the concept of point are physically possible. This will be discussed in the third part of the adventure.

Challenge 108, page 69: One can rotate the hand in a way that the arm makes the motion described here. See also page 731.

Challenge 109, page 69: Any number, without limit.

Challenge 110, page 69: The blood and nerve supply would periodically be run over by the wheel. Worse, if the wheel has an axis, the method cannot be applied! It thus becomes impossible to make a wheel *axis* using a single piece of skin. (Even if blood supply technologies like continuous flow reactors were used, animals could not make such a detached wheel grow in a way tuned to the rest of the body and they would have difficulties repairing a damaged wheel. Wheels cannot be grown on animals.) Could such a connection realize a propeller?

Challenge 111, page 70: The brain in the skull, the blood factories inside bones or the growth of the eye are examples.

Challenge 112, page 71: One can also add the Sun, the sky and the landscape to the list.

Challenge 113, page 71: Ghosts, hallucinations, Elvis sightings, or extraterrestrials must all be one or the other. There is no third option. Even shadows are only special images.

Challenge 114, page 71: The issue was hotly discussed in the seventeenth century; even Galileo argued for them being images. However, they are objects, as they can collide with other objects, as the spectacular comet-Jupiter collision in 1994 showed.

Challenge 115, page 72: The minimum speed is roughly the one at which it is possible to ride without hands. If you do so, and then *gently* push on the steering wheel, you can make the experience described above. Watch out: too strong a push will make you fall badly.

Challenge 116, page 74: If the ball is not rotating, after the collision the two balls will depart with a *right* angle between them.

Challenge 117, page 74: Part of the energy is converted into heat; the rest is transferred as kinetic energy of the concrete block. As the block is heavy, its speed is small and easily stopped by the human body. This effect works also with anvils, it seems. In another common variation the person does not lie on nails, but on air: he just keeps himself horizontal, with head and shoulders on one chair, and the feet on a second one.

Challenge 118, page 75: Yes, mass works also for magnetism, because the precise condition is not that the interaction be central, but that it realizes a more general condition, which includes accelerations such as those produced by magnetism. Can you deduce the condition from the definition of mass?

Challenge 119, page 76: The weight decreased due to the evaporated water lost by sweating and, to a minor degree, due to the exhaled carbon bound in carbon dioxide.

Challenge 120, page 76: Rather than using the inertial effects of the Earth, it is easier to deduce its mass from its gravitational effects.

Challenge 124, page 77: At first sight, relativity implies that tachyons have imaginary mass; however, the imaginary factor can be extracted from the mass–energy and mass–momentum relation, so that one can define a real mass value for tachyons; as a result, faster tachyons have smaller energy and smaller momentum. Both can be a negative number of any size.

Challenge 125, page 77: Legs are never perfectly vertical; they would immediately glide away. Once the cat or the person is on the floor, it is almost impossible to stand up again.

Challenge 126, page 77: Momentum (or centre of mass) conservation would imply that the environment would be accelerated into the opposite direction. Energy conservation would imply that a huge amount of energy would be transferred between the two locations, melting everything in between. Teleportation would thus contradict energy and momentum conservation.

Challenge 127, page 78: The part of the tides due to the Sun, the solar wind, and the interactions between both magnetic fields are examples of friction mechanisms.

Challenge 128, page 79: With the factor $1/2$, increase of (physical) kinetic energy is equal to the (physical) work performed on a system.

Challenge 130, page 79: It is a smart application of momentum conservation.

Challenge 131, page 80: Neither. With brake on, the damage is higher, but still equal for both cars.

Challenge 132, page 81: Heating systems, transport engines, engines in factories, steel plants, electricity generators covering the losses in the power grid, etc.

Challenge 136, page 84: Just throw it into the air and compare your effort to make it turn around various axes.

Challenge 137, page 84: Use the definition of the moment of inertia and Pythagoras' theorem for every mass element of the body.

Challenge 138, page 85: Hang up the body, attaching the rope in two different points. The crossing point of the prolonged rope lines is the centre of mass.

Challenge 139, page 85: Spheres have an orientation, because we can always add a tiny spot on their surface. This possibility is not given for microscopic objects, and we shall study this situation in the part on quantum theory.

Challenge 142, page 85: See Tables 13 and 14.

Challenge 143, page 86: Self-propelled linear motion contradicts the conservation of momentum; self-propelled rotation does not. But the deep, final reason for the difference will be unveiled in the third part of our adventure.

Challenge 144, page 86: Yes, the ape can reach the banana. The ape just has to turn around its own axis. For every turn, the plate will rotate a bit towards the banana. Of course, other methods, like blowing at a right angle to the axis, peeing, etc., are also possible.

Challenge 146, page 87: The points that move exactly along the radial direction of the wheel form a circle below the axis. They are the points that are sharp in Figure 37 of page 87.

Challenge 147, page 87: Use the conservation of angular momentum around the point of contact. If all the wheel's mass is assumed in the rim, the final rotation speed is half the initial one; it is independent of the friction coefficient.

Challenge 151, page 91: A short pendulum of length L that swings in two dimensions (with amplitude ρ and orientation φ) shows two additional terms in the Lagrangian $\mathcal{L}$:

$$\mathcal{L} = T - V = \frac{1}{2}m\dot{\rho}^2(1+\frac{\rho^2}{L^2}) + \frac{l_z^2}{2m\rho^2} - \frac{1}{2}m\omega_0^2\rho^2(1+\frac{\rho^2}{4\,L^2}) \tag{838}$$

where as usual the basic frequency is $\omega_0^2 = g/L$ and the angular momentum is $l_z = m\rho^2\dot{\varphi}$. The two additional terms disappear when $L \to \infty$; in that case, if the system oscillates in an ellipse with semiaxes a and b, the ellipse is fixed in space, and the frequency is ω_0. For *finite* pendulum length L, the frequency changes to

$$\omega = \omega_0(1 - \frac{a^2+b^2}{16\,L^2}) \quad ; \tag{839}$$

most of all, the ellipse turns with a frequency

$$\Omega = \omega\frac{3}{8}\frac{ab}{L^2} \; . \tag{840}$$

(These formulae can be derived using the least action principle, as shown by C.G. Gray, G. Karl & V.A. Novikov, Progress in classical and quantum variational principles, available as preprint at http://www.arxiv.org/abs/physics/0312071.) In other words, a short pendulum in elliptical motion shows a precession even *without* the Coriolis effect. Since this precession frequency diminishes with $1/L^2$, the effect is small for long pendulums, where only the Coriolis effect is left over. To see the Coriolis effect in a short pendulum, one thus has to avoid that it starts swinging in an elliptical orbit by adding a suppression method of elliptical motion.

Challenge 152, page 92: The Coriolis acceleration is the reason for the deviation from the straight line. The Coriolis acceleration is due to the change of speed with distance from the rotation axis. Now think about a pendulum, located in Paris, swinging in the North-South direction. At the Southern end of the swing, the pendulum is further from the axis by $A \sin \varphi$, where A is the swing amplitude. At that end of the swing, the central support point overtakes the pendulum bob with a relative horizontal speed given by $v = 2\pi A \sin \varphi / 24\,\mathrm{h}$. The period of precession is given by $T_{\mathrm{F}} = v/2\pi A$, where $2\pi A$ is the circumference $2\pi A$ of the envelope of the pendulum's path (relative to the Earth). This yields $T_{\mathrm{F}} = 24\,\mathrm{h}/\sin \varphi$.

Challenge 153, page 92: Rotation leads to a small frequency and thus colour changes of the circulating light.

Challenge 154, page 92: The weight changes when going east or when moving west due to the Coriolis acceleration. If the rotation speed is tuned to the oscillation frequency of the balance, the effect is increased by resonance. This trick was also used by Eötvös.

Challenge 155, page 92: The Coriolis acceleration makes the bar turn, as every moving body is deflected to the side, and the two deflections add up in this case. The direction of the deflection depends on whether the experiments is performed on the northern or the southern hemisphere.

Challenge 156, page 92: When rotated by π around an east–west axis, the Coriolis force produces a drift velocity of the liquid around the tube. It has the value

$$v = 2\omega r \sin \theta ,$$

as long as friction is negligible. Here ω is the angular velocity of the Earth, θ the latitude and r the (larger) radius of the torus. For a tube with 1 m diameter in continental Europe, this gives a speed of about $6.3 \cdot 10^{-5}\,\mathrm{m/s}$.

The measurement can be made easier if the tube is restricted in diameter at one spot, so that the velocity is increased there. A restriction by an area factor of 100 increases the speed by the same factor. When the experiment is performed, one has to carefully avoid any other effects that lead to moving water, such as temperature gradients across the system.

Challenge 158, page 97: The original result by Bessel was $0.3136''$, or 657.7 thousand orbital radii, which he thought to be 10.3 light years or 97.5 Pm.

Challenge 160, page 99: The galaxy forms a stripe in the sky. The galaxy is thus a flattened structure. This is even clearer in the infrared, as shown more clearly in Figure 190 on page 403. From the flattening (and its circular symmetry) we can deduce that the galaxy must be rotating. Thus other matter must exist in the universe.

Challenge 134, page 83: If the Earth changed its rotation speed ever so slightly we would walk inclined, the water of the oceans would flow north, the atmosphere would be filled with storms and earthquakes would appear due to the change in Earth's shape.

Challenge 163, page 100: The scale reacts to your heartbeat. The weight is almost constant over time, except when the heart beats: for a short duration of time, the weight is somewhat lowered at each beat. Apparently it is due to the blood hitting the aortic arch when the heart pumps it upwards. The speed of the blood is about 0.3 m/s at the maximum contraction of the left ventricle. The distance to the aortic arch is a few centimetre. The time between the contraction and the reversal of direction is about 15 ms.

Challenge 161, page 100: Probably the 'rest of the universe' was meant by the writer. Indeed, a moving a part never shifts the centre of gravity of a closed system. But is the universe closed? Or a system? The third part of the adventure centres on these issues.

Challenge 164, page 100: The conservation of angular momentum saves the glass. Try it.

Challenge 166, page 100: Assuming a square mountain, the height h above the surrounding crust and the depth d below are related by

$$\frac{h}{d} = \frac{\rho_{\mathrm{m}} - \rho_{\mathrm{c}}}{\rho_{\mathrm{c}}} \tag{841}$$

where ρ_{c} is the density of the crust and ρ_{m} is the density of the mantle. For the density values given, the ratio is 6.7, leading to an additional depth of 6.7 km below the mountain.

Challenge 170, page 101: The behaviour of the spheres can only be explained by noting that elastic waves propagate through the chain of balls. Only the propagation of these elastic waves, in particular their reflection at the end of the chain, explains that the same number of balls that hit on one side are lifted up on the other. For long times, friction makes all spheres oscillate in phase. Can you confirm this?

Challenge 171, page 102: When the short cylinder hits the long one, two compression waves start to run from the point of contact through the two cylinders. When each compression wave arrives at the end, it is reflected as an expansion wave. If the geometry is well chosen, the expansion wave coming back from the short cylinder can continue into the long one (which is still in his compression phase). In practice, length ratios of 1:10 are used in hammer drills. The waves from the short cylinder can thus depose much of their energy into the long cylinder. Momentum is conserved, as is energy; the long cylinder is oscillating in length when it detaches, so that not all its energy is kinetic energy.

Challenge 172, page 102: The momentum transfer to the wall is double when the ball rebounds perfectly.

Challenge 173, page 102: If the cork is in its intended position: take the plastic cover off the cork, put the cloth around the bottle (this is for protection reasons only) and repeatedly hit the bottle on the floor or a fall in an inclined way, as shown in Figure 32 on page 77. With each hit, the cork will come out a bit.

If the cork has fallen inside the bottle: put half the cloth inside the bottle; shake until the cork falls unto the cloth. Pull the cloth out: first slowly, until the cloth almost surround the cork, and then strongly.

Challenge 175, page 102: The atomic force microscope.

Challenge 176, page 103: Use Figure 41 on page 90 for the second half of the trajectory, and think carefully about the first half.

Challenge 177, page 103: Starting rockets at the Equator saves a lot of energy, thus of fuel and of weight.

Challenge 178, page 103: Running man: $E \approx 0.5 \cdot 80\,\mathrm{kg} \cdot (5\,\mathrm{m/s})^2 = 1\,\mathrm{kJ}$; rifle bullet: $E \approx 0.5 \cdot 0.04\,\mathrm{kg} \cdot (500\,\mathrm{m/s})^2 = 5\,\mathrm{kJ}$.

Challenge 179, page 103: The flame leans towards the inside.

Challenge 180, page 103: The ball leans in the direction it is accelerated to. Yes, one could imagine that the ball in a glass at rest pulls upwards because the floor is accelerated upwards. We will come back to this issue in the section of general relativity.

Challenge 181, page 103: It almost doubles in size.

Challenge 184, page 103: Place the tea in cups on a board and attach the board to four long ropes that you keep in your hand.

Challenge 188, page 104: An earthquake with Richter magnitude of 12 is 1000 times the energy of the 1960 Chile quake with magnitude 10; the latter was due to a crack throughout the full 40 km of the Earth's crust along a length of 1000 km in which both sides slipped by 10 m with respect to each other. Only the impact of a meteorite could lead to larger values than 12.

Challenge 189, page 104: This is not easy; a combination of friction and torques play a role. See for example the article ...

Challenge 191, page 104: If a wedding ring rotates on an axis that is not a principal one, angular momentum and velocity are not parallel.

Challenge 192, page 105: Yes; it happens twice a year. To minimize the damage, dishes should be dark in colour.

Challenge 193, page 105: A rocket fired from the back would be a perfect defence against planes attacking from behind. However, when released, the rocket is effectively flying backwards with respect to the air, thus turns around and then becomes a danger to the plane that launched it. Engineers who did not think about this effect almost killed a pilot during the first such tests.

Challenge 195, page 105: Whatever the ape does, whether it climbs up or down or even lets himself fall, it remains at the same height as the mass. What happens with friction at the wheel?

Challenge 197, page 105: Weigh the bullet and shoot it against a mass hanging from the ceiling. From the mass and the angle it is deflected to, the momentum of the bullet can be determined.

Challenge 199, page 105: Yes, if he moves at a large enough angle to the direction of the boat's motion.

Challenge 201, page 105: $\Theta = \frac{2}{5}mr^2$.

Challenge 203, page 105: See the article by C. Ucke & H.-J. Schlichting, Faszinierendes Dynabee, *Physik in unserer Zeit* 33, pp. 230–231, 2002.

Challenge 204, page 105: Yes. Can you imagine what happens for an observer on the Equator?

Challenge 205, page 105: A straight line at the zenith, and circles getting smaller at both sides. See an example on the website http://antwrp.gsfc.nasa.gov/apod/ap021115.html.

Challenge 207, page 106: The plane is described in the websites cited; for a standing human the plane is the vertical plane containing the two eyes.

Challenge 208, page 107: As said before, legs are simpler than wheels to grow, to maintain and to repair; in addition, legs do not require flat surfaces (so-called 'streets') to work.

Challenge 210, page 108: Classical or everyday nature is right-left symmetric and thus requires an even number of legs. Walking on two-dimensional surfaces naturally leads to a minimum of four legs.

Challenge 212, page 108: The length of the day changes with latitude. So does the length of a shadow or the elevation of stars at night, facts that are easily checked by telephoning a friend. Ships appear at the horizon first be showing only their masts.

Challenge 213, page 108: Robert Peary had forgotten that on the date he claimed to be at the North Pole, 6th of April 1909, the Sun is very low on the horizon, casting very long shadows, about ten times the height of objects. But on his photograph the shadows are much shorter. (In fact, the picture is taken in such a way to hide all shadows as carefully as possible.) Interestingly, he had even convinced the US congress to officially declare him the first man on the North Pole in 1911. (A rival had claimed to have reached it earlier on, but his photograph has the same mistake.)

Challenge 214, page 108: Yes, the effect has been measured for skyscrapers. Can you estimate the values?

Challenge 215, page 110: The tip of the velocity arrow, when drawn over time, produces a circle around the centre of motion.

Challenge 218, page 110: The value of the product GM for the Earth is $4.0 \cdot 10^{14}\ \mathrm{m^3/s^2}$.

Challenge 219, page 110: All points can be reached; but when shooting horizontally in one given direction, only points on the first half of the circumference can be reached.

Challenge 222, page 111: The Atwood machine is the answer: two almost equal weights connected by a string hanging from a well-oiled wheel. The heavier one falls very slowly. Can you determine the acceleration as a function of the two masses?

Challenge 224, page 112: The speed is proportional to l/T, which makes it proportional to $l^{1/2}$.

Challenge 226, page 112: Cavendish suspended a horizontal handle with a long metal wire. He then approached a large mass to the handle, avoiding any air currents, and measured how much the handle rotated.

Challenge 227, page 113: The acceleration due to gravity is $a = Gm/r^2 \approx 5\,\mathrm{nm/s^2}$ for a mass of 75 kg. For a fly with mass $m_{\mathrm{fly}} = 0.1\,\mathrm{g}$ landing on a person with a speed of $v_{\mathrm{fly}} = 1\,\mathrm{cm/s}$ and deforming the skin (without energy loss) by $d = 0.3\,\mathrm{mm}$, a person would be accelerated by $a = (v^2/d)(m_{\mathrm{fly}}/m) = 0.4\,\mathrm{\mu m/s^2}$. The energy loss of the inelastic collision reduces this value at least by a factor of ten.

Challenge 234, page 117: A flash of light is sent to the Moon, where a Cat's-eye has been deposited by the cosmonauts in 1969. The measurement precision of the time a flash take to go and come back is sufficient to measure the Moon's distance change.

Challenge 240, page 120: This is a resonance effect, in the same way that a small vibration of a string can lead to large oscillation of the air and sound box in a guitar.

Challenge 242, page 123: The total angular momentum of the Earth and the Moon must remain constant.

Challenge 248, page 125: The centre of mass falls with the usual acceleration; the end thus falls faster.

Challenge 249, page 126: Just use energy conservation for the two masses of the jumper and the string. For more details, including the comparison of experimental measurements and theory, see N. Dubelaar & R. Brantjes, De valversnelling bij bungee-jumping, *nederlands tijdschrift voor natuurkunde* **69**, pp. 316–318, October 2003.

Challenge 250, page 126: About 1 ton.

Challenge 251, page 126: About 5 g.

Challenge 252, page 126: Your weight is roughly constant; thus the Earth must be round. On a flat Earth, the weight would change from place to place.

Challenge 253, page 126: Nobody ever claimed that the center of mass is the same as the centre of gravity! The attraction of the Moon is negligible on the surface of the Earth.

Challenge 255, page 126: That is the mass of the Earth. Just turn the table on its head.

Challenge 258, page 127: The Moon will be about 1.25 times as far as it is now. The Sun then will slow down the Earth–Moon system rotation, this time due to the much smaller tidal friction from the Sun's deformation. As a result, the Moon will return to smaller and smaller distances to Earth. However, the Sun will have become a red giant by then, after having swallowed both the Earth and the Moon.

Challenge 260, page 127: As Galileo determined, for a swing (half a period) the ratio is $\sqrt{2}/pi$. Not more than two, maybe three decimals of π can be determined this way.

Challenge 261, page 127: Momentum conservation is not a hindrance, as any tennis racket has the same effect on the tennis ball.

Challenge 262, page 127: In fact, in velocity space, elliptic, parabolic and hyperbolic motions are all described by circles. In all cases, the hodograph is a circle. Ref. 95

Challenge 263, page 127: This question is old (it was already asked in Newton's times) and deep. One reason is that stars are kept apart by rotation around the galaxy. The other is that galaxies are kept apart by the momentum they got in the big bang. Without the big bang, all stars would have collapsed together. In this sense, the big bang can be deduced from the attraction of gravitation and the immobile sky at night. We shall find out later that the darkness of the night sky gives a second argument for the big bang.

Challenge 264, page 127: Due to the plateau, the effective mass of the Earth is larger.

Challenge 265, page 127: The choice is clear once you notice that there is no section of the orbit which is concave towards the Sun. Can you show this?

Challenge 266, page 128: It would be a black hole; no light could escape. Black holes are discussed in detail in the chapter on general relativity.

Challenge 268, page 128: A handle of two bodies.

Challenge 270, page 128: Using a maximal jumping height of h =0.5 m on Earth and an estimated asteroid density of ρ =3 Mg/m^3, we get a maximum radius of $R^2 = 3gh/4\pi G\rho \approx 703$ m.

Challenge 274, page 130: The tunnel would be an elongated ellipse in the plane of the Equator, reaching from one point of the Equator to the point at the antipodes. The time of revolution would not change, compared to a non-rotating Earth. See A.J. Simonson, Falling down a hole through the Earth, *Mathematics Magazine* 77, pp. 171–188, June 2004.

Challenge 277, page 130: First, during northern summer time the Earth moves faster around the Sun than during northern winter time. Second, shallow Sun's orbits on the sky give longer days because of light from when the Sun is below the horizon.

Challenge 281, page 130: True.

Challenge 285, page 131: Never. The Moon points always towards the Earth. The Earth changes position a bit, due to the ellipticity of the Moon's orbit. Obviously, the Earth shows phases.

Challenge 287, page 131: What counts is local verticality; with respect to it, the river always flows down.

Challenge 288, page 131: There are no such bodies, as the chapter of general relativity will show.

Challenge 290, page 132: The oscillation is a purely sinusoidal, or harmonic oscillation, as the restoring force increases linearly with distance from the centre of the Earth. The period T for a homogeneous Earth is $T = 2\pi\sqrt{R^3/GM} = 84$ min. The same result holds for all tunnels, even if the tunnel does not go through the centre of the Earth. See for example, R.H. Romer, The answer is forty-two – many mechanics problems, only one answer, *Physics Teacher* **41**, pp. 286–290, May 2003.

Challenge 291, page 132: The period is the same for all such tunnels and thus in particular it is the same as the 84 min valid also for the pole to pole tunnel. See challenge 290.

Challenge 295, page 135: The electricity consumption of a rising escalator indeed increases when the person on it walks upwards. By how much?

Challenge 296, page 135: Knowledge is power. Time is money. Now, power is defined as work per time. Inserting the previous equations and transforming them yields

$$\text{money} = \frac{\text{work}}{\text{knowledge}} , \qquad (842)$$

which shows that the less you know, the more money you make. That is why scientists have low salaries.

Challenge 300, page 138: True?

Challenge 303, page 138: From $\mathrm{d}v/\mathrm{d}t = g - v^2(1/2c_w A\rho/m)$ and using the abbreviation $c = 1/2c_w A\rho$, we can solve for $v(t)$ by putting all terms containing the variable v on one side, all terms with t on the other, and integrating on both sides. We get $v(t) = \sqrt{gm/c}\tanh\sqrt{cg/m}\,t$.

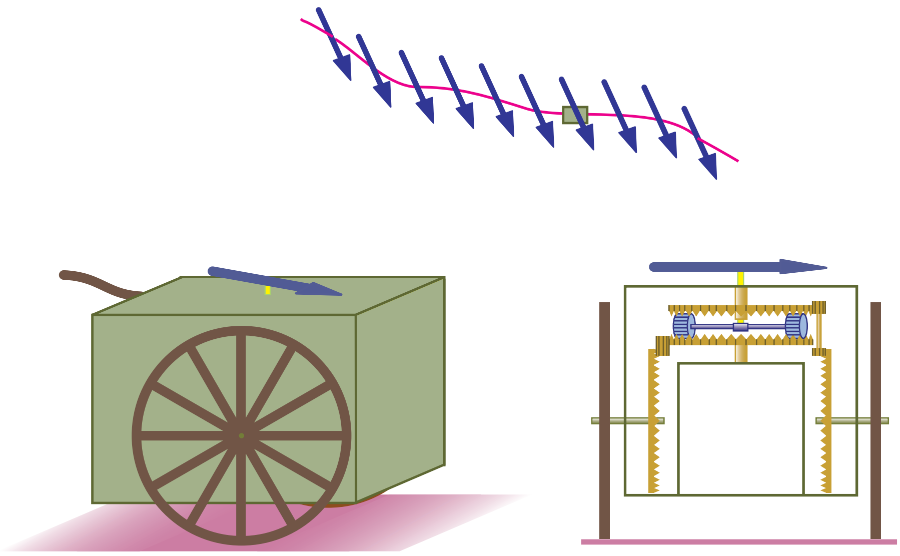


Figure 385 The south-pointing carriage

Page 529 **Challenge 306**, page 140: The light mill is an example.

Challenge 307, page 140: Electric charge.

Challenge 308, page 140: If you have found reasons to answer yes, you overlooked something. Just go into more details and check whether the concepts you used apply to the universe. Also define carefully what you mean by 'universe'.

Challenge 310, page 142: A system showing energy or matter motion faster than light would imply that for such systems there are observers for which the order between cause and effect are reversed. A space-time diagram (and a bit of exercise from the section on special relativity) shows this.

Challenge 311, page 142: This connection shall become important in the third part of our adventure.

Challenge 315, page 144: Of course; moral laws are summaries of what others think or will do about personal actions.

Challenge 316, page 144: Space-time is defined using matter; matter is defined using space-time.

Challenge 317, page 144: Fact is that physics has been based on a circular definition for hundreds of years. Thus it is possible to build even an exact science on sand. Nevertheless, the elimination of the circularity is an important aim.

Challenge 318, page 148: For example, speed inside materials is slowed, but between atoms, light still travels with vacuum speed.

Challenge 321, page 159: Figure 385 shoes the most credible reconstruction of a south-pointing carriage.

Challenge 322, page 160: The water is drawn up along the sides of the spinning egg. The fastest way to empty a bottle of water is to spin the water while emptying it.

Challenge 323, page 160: The right way is the one where the chimney falls like a V, not like an inverted V. See challenge 248 on falling brooms for inspiration on how to deduce the answer.

Challenge 331, page 166: In one dimension, the expression $F = ma$ can be written as $-\mathrm{d}V/\mathrm{d}x = m\mathrm{d}^2x/\mathrm{d}t^2$. This can be rewritten as $\mathrm{d}(-V)/\mathrm{d}x - \mathrm{d}/\mathrm{d}t[\mathrm{d}/\mathrm{d}\dot{x}(\frac{1}{2}m\dot{x}^2)] = 0$. This can be expanded to $\partial/\partial x(\frac{1}{2}m\dot{x}^2 - V(x)) - \mathrm{d}/[\partial/\partial\dot{x}(\frac{1}{2}m\dot{x}^2 - V(x))] = 0$, which is Lagrange's equation for this case.

Challenge 333, page 166: Do not despair. Up to now, nobody has been able to imagine a universe (that is not necessarily the same as a 'world') different from the one we know. So far, such attempts have always led to logical inconsistencies.

Challenge 335, page 167: The two are equivalent since the equations of motion follow from the principle of minimum action and at the same time the principle of minimum action follows from the equations of motion.

Challenge 337, page 168: For gravity, all three systems exist: rotation in galaxies, pressure in planets and the Pauli pressure in stars. Against the strong interaction, the Pauli principle acts in nuclei and neutron stars; in neutron stars maybe also rotation and pressure complement the Pauli pressure. But for the electromagnetic interaction there are no composites other than our everyday matter, which is organized by the Pauli principle alone.

Challenge 339, page 171: Angular momentum is the change with respect to angle, whereas rotational energy is again the change with respect to time, as all energy is.

Challenge 340, page 171: Not in this way. A small change can have a large effect, as every switch shows. But a small change in the brain must be communicated outside, and that will happen roughly with a $1/r^2$ dependence. That makes the effects so small, that even with the most sensitive switches – which for thoughts do not exist anyway – no effects can be realized.

Challenge 344, page 172: The relation is

$$\frac{c_1}{c_2} = \frac{\sin\alpha_1}{\alpha_2} . \tag{843}$$

The particular speed ratio between air (or vacuum, which is almost the same) and a material gives the *index of refraction* n:

$$n = \frac{c_1}{c_0} = \frac{\sin\alpha_1}{\alpha_0} \tag{844}$$

Challenge 345, page 172: Gases are mainly made of vacuum. Their index of refraction is near to one.

Challenge 346, page 172: Diamonds also sparkle because they work as prisms; different colours have different indices of refraction. Thus their sparkle is also due to their dispersion; therefore it is a mix of all colours of the rainbow.

Challenge 347, page 172: The principle for the growth of trees is simply the minimum of potential energy, since the kinetic energy is negligible. The growth of vessels inside animal bodies is minimized for transport energy; that is again a minimum principle. The refraction of light is the path of shortest time; thus it minimizes change as well, if we imagine light as moving entities moving without any potential energy involved.

Challenge 348, page 173: Special relativity requires that an invariant measure of the action exist. It is presented later in the walk.

Challenge 349, page 173: The universe is not a physical system. This issue will be discussed in detail later on. Page 966

Challenge 350, page 173: We talk to a person because we know that somebody understands us. Thus we assume that she somehow sees the same things we do. That means that observation is partly viewpoint-independent. Thus nature is symmetric.

Challenge 351, page 174: Memory works because we recognize situations. This is possible because situations over time are similar. Memory would not have evolved without this reproducibility.

Challenge 353, page 176: The integers under addition form a group. Does a painter's set of oil colours with the operation of mixing form a group?

Challenge 360, page 180: Scalar is the magnitude of any vector; thus the speed, defined as $v = |\mathbf{v}|$, is a scalar, whereas the velocity $\mathbf{v}$ is not. Thus the length of any vector (or pseudovector), such as force, acceleration, magnetic field, or electric field, is a scalar, whereas the vector itself is not a scalar.

Challenge 363, page 181: The charge distribution of an extended body can be seen as a sum of a charge, a charge dipole, a charge quadrupole, a charge octupole, etc. The quadrupole is described by a tensor.

Compare: The inertia against motion of an extended body can be seen as sum of a mass, a mass dipole, a mass quadrupole, a mass octupole, etc. The mass quadrupole is described by the moment of inertia.

Challenge 367, page 183: The conserved charge for rotation invariance is angular momentum.

Challenge 370, page 187: An oscillation has a period in time, i.e. a discrete time translation symmetry. A wave has both discrete time and discrete space translation symmetry.

Challenge 371, page 187: Motion reversal is a symmetry for any closed system; despite the observations of daily life, the statements of thermodynamics and the opinion of several famous physicists (who form a minority though) all ideally closed systems are reversible.

Challenge 379, page 191: The potential energy is due to the 'bending' of the medium; a simple displacement produces no bending and thus contains no energy. Only the gradient captures the bending idea.

Challenge 382, page 192: Waves can be damped to extremely low intensities. If this is not possible, the observation is not a wave.

Challenge 383, page 193: Page 518 tells how to observe diffraction and interference with your naked fingers.

Challenge 392, page 198: The sound of thunder or of car traffic gets lower and lower in frequency with increasing distance.

Challenge 394, page 198: Neither; both possibilities are against the properties of water: in surface waves, the water molecules move in circles.

Challenge 395, page 199: Swimmers are able to cover 100 m in 48 s, or slightly better than 2 m/s. With a body length of about 1.9 m, the critical speed is 1.7 m/s. That is why short distance swimming depends on training; for longer distances the technique plays a larger role, as the critical speed has not been attained yet. The formula also predicts that on the 1500 m distance, a 2 m tall swimmer has a potential advantage of over 45 s on one with body height of 1.8 m. In addition, longer swimmers have an additional advantage: they swim shorter distances (why?). It is thus predicted that successful long-distance swimmers will get taller and taller over time. This is a pity for a sport that so far could claim to have had champions of all sizes and body shapes, in contrast to many other sports.

Challenge 397, page 199: To reduce noise reflection and thus hall effects. They effectively diffuse the arriving wave fronts.

Challenge 399, page 200: Waves in a river are never elliptical; they remain circular.

Challenge 400, page 200: The lens is a cushion of material that is 'transparent' to sound. The speed of sound is faster in the cushion than in the air, in contrast to a glass lens, where the speed of light is slower in the glass. The shape is thus different: the cushion must look like a biconcave lens.

Challenge 402, page 200: The Sun is always at a different position than the one we observe it to be. What is the difference, measured in angular diameters of the Sun?

Challenge 403, page 200: The $3 \times 3 \times 3$ cube has a rigid system of three perpendicular axes, on which a square can rotate at each of the 6 ends. The other squares are attaches to pieces moving around theses axes. The $4 \times 4 \times 4$ cube is different though; just find out. The limit on the segment number seems to be 6, so far. A $7 \times 7 \times 7$ cube requires varying shapes for the segments. But more than $5 \times 5 \times 5$ is not found in shops. However, the website http://www.oinkleburger.com/Cube/applet/ allows to play with virtual cubes up to $100 \times 100 \times 100$ and more.

Challenge 405, page 200: An overview of systems being tested at present can be found in K.-U. Graw, Energiereservoir Ozean, *Physik in unserer Zeit* **33**, pp. 82–88, Februar 2002. See also Oceans of electricity – new technologies convert the motion of waves into watts, *Science News* **159**, pp. 234–236, April 2001.

Challenge 406, page 200: In everyday life, the assumption is usually justified, since each spot can be approximately represented by an atom, and atoms can be followed. The assumption is questionable in situations such as turbulence, where not all spots can be assigned to atoms, and most of all, in the case of motion of the vacuum itself. In other words, for gravity waves, and in particular for the quantum theory of gravity waves, the assumption is not justified.

Challenge 408, page 202: There are many. One would be that the transmission and thus reflection coefficient for waves would almost be independent of wavelength.

Challenge 409, page 202: A drop with a diameter of 3 mm would cover a surface of 7.1 m^2 with a 2 nm film.

Challenge 410, page 203: For jumps of an animal of mass m the necessary energy E is given as $E = mgh$, and the work available to a muscle is roughly speaking proportional to its mass $W \sim m$. Thus one gets that the height h is independent of the mass of the animal. In other words, the specific mechanical energyof animals is around 1.5 ± 0.7 J/kg.

Challenge 412, page 204: The critical height for a column of material is given by $h_{\text{crit}}^4 = \frac{\beta}{4\pi g} m \frac{E}{\rho^2}$, where $\beta \approx 1.9$ is the constant determined by the calculation when a column buckles und its own weight.

Challenge 414, page 208: Throwing the stone makes the level fall, throwing the water or the piece of wood leaves it unchanged.

Challenge 415, page 208: No metal wire allows to build such a long wire. Only the idea of carbon nanotubes has raised the hope again; some dream of wire material based on them, stronger than any material known so far. However, no such material is known yet. The system faces many dangers, such as fabrication defects, lightning, storms, meteorites and space debris. All would lead to the breaking of the wires – if such wires will ever exist. But the biggest of all dangers is the lack of cash to build it.

Challenge 421, page 208: This argument is comprehensible only when one remembers that 'twice the amount' means 'twice as many molecules'.

Challenge 422, page 208: The alcohol is frozen and the chocolate is put around it.

Challenge 423, page 209: Building such a machine is envisaged by at least one researcher. I suggest that is should be based on the same machines that throw the clay pigeons used in the sports of trap shooting and skeet.

Challenge 424, page 209: The third component of *air* is the noble gas argon, making up about 1 %. The rest is made up by carbon dioxide, water vapour and other gases. Are these percentages volume or weight percentages?

Challenge 425, page 209: It uses the air pressure created by the water flowing downwards.

Challenge 426, page 209: Yes. The bulb will not resist two such cars though.

Challenge 429, page 209: None.

Challenge 430, page 209: He brought the ropes into the cabin by passing them through liquid mercury.

Challenge 431, page 209: The pressure destroys the lung.

Challenge 433, page 210: Either they fell on inclined snowy mountain sides, or they fell into high trees, or other soft structures. The record was over 7 km of survived free fall.

Challenge 434, page 210: The blood pressure in the feet of a standing human is about 27 kPa, double the pressure at the heart.

Challenge 435, page 210: Calculation gives $N = J/j = 0.0001\,\mathrm{m^3/s}/(7\,\mathrm{\mu m^2}\,0.0005\,\mathrm{m/s})$, or about $6 \cdot 10^9$; in reality, the number is much larger, as most capillaries are closed at a given instant. The reddening of the face shows what happens when all small blood vessels are opened at the same time.

Challenge 436, page 210: The soap flows down the bulb, making it thicker at the bottom and thinner at the top, until it bursts.

Challenge 437, page 210: A medium-large earthquake would be generated.

Challenge 438, page 210: A stalactite contains a thin channel along its axis through which the water flows, whereas a stalagmite is massive throughout.

Challenge 440, page 210: About 1 part in a thousand.

Challenge 441, page 210: For this to happen, friction would have to exist on the microscopic scale and energy would have to disappear.

Challenge 442, page 210: The longer funnel is empty before the short one. Energy conservation yields $P/\rho + gh + v^2/2 = \text{const}$. Thus the speed v is higher for greater heights h of the funnel.

Challenge 443, page 210: The eyes of fish are positioned in such a way that the pressure reduction by the flow is compensated by the pressure increase of the stall. By the way, their heart is positioned in such a way that it is helped by the underpressure.

Challenge 445, page 211: Glass shatters, glass is elastic, glass shows transverse sound waves, glass does not flow (in contrast to what many books state), not even on scale of centuries, glass molecules are fixed in space, glass is crystalline at small distances, a glass pane supported at the ends does not hang through.

Challenge 446, page 211: This feat has been achieved for lower mountains, such as the Monte Bianco in the Alps. At present however, there is no way to safely hover at the high altitudes of the Himalayas.

Challenge 448, page 211: The iron core of the Earth formed in the way described by collecting the iron from colliding asteroids. However, the Earth was more liquid at that time. The iron will most probably not sink. In addition, there is no known way to make the measurement probe described.

Challenge 449, page 211: Press the handkerchief in the glass, and lower the glass into the water with the opening first, while keeping the opening horizontal. This method is also used to lower people below the sea. The paper ball in the bottle will fly towards you. Blowing into a funnel will keep the ping-pong ball tightly into place, and the more so the stronger you blow. Blowing through a funnel towards a candle will make it lean towards you.

Challenge 452, page 216: In 5000 million years, the present method will stop, and the Sun will become a red giant. abut it will burn for many more years after that.

Challenge 459, page 221: The answer depends on the size of the balloons, as the pressure is not a monotonous function of the size. If the smaller balloon is not too small, the smaller balloon wins.

Challenge 460, page 221: Measure the area of contact between tires and street (all four) and then multiply by 200 kPa, the usual tire pressure. You get the weight of the car.

Challenge 464, page 223: If the average square displacement is proportional to time, the matter is made of smallest particles. This was confirmed by the experiments of Jean Perrin. The next step is to deduce the number number of these particles form the proportionality constant. This constant, defined by $\langle d^2 \rangle = 4Dt$, is called the diffusion constant (the factor 4 is valid for random motion in two dimensions). The diffusion constant can be determined by watching the motion of a particle under the microscope.

We study a Brownian particle of radius a. In two dimensions, its square displacement is given by

$$\langle d^2 \rangle \frac{4kT}{\mu} t \; , \qquad (845)$$

where k is the Boltzmann constant and T the temperature. The relation is deduced by studying the motion of a particle with drag force $-\mu v$ that is subject to random hits. The linear drag coefficient μ of a sphere of radius a is given by

$$\mu = 6\pi\eta a \; . \qquad (846)$$

In other words, one has

$$k = \frac{6\pi\eta a}{4T} \frac{\langle d^2 \rangle}{t} \; . \qquad (847)$$

All quantities on the right can be measured, thus allowing to determine the Boltzmann constant k. Since the ideal gas relation shows that the ideal gas constant R is related to the Boltzmann constant by $R = N_A k$, the Avogadro constant N_A that gives the number of molecules in a mole is also found in this way.

Challenge 456, page 220: We will find out later that the universe is not a system; thus the concept of entropy does not apply to it. Thus the universe is neither isolated nor closed. Page 966

Challenge 471, page 228: Yes, the effect is easily noticeable.

Challenge 473, page 229: Hot air is less dense and thus wants to rise.

Challenge 475, page 229: The air had to be dry.

Challenge 476, page 229: In general, it is impossible to draw a line through three points.

Challenge 477, page 229: No, as a water molecule is heavier than that. However, if the water is allowed to be dirty, it is possible. What happens if the quantum of action is taken into account?

Challenge 478, page 229: Keep the paper wet.

Challenge 479, page 229: The danger is not due to the amount of energy, but due to the time in which it is available.

Challenge 480, page 230: The internet is full of solutions.

Challenge 482, page 230: Only if it is a closed system. Is the universe closed? Is it a system? This is discussed in the third part of the mountain ascent.

Challenge 485, page 230: For such small animals the body temperature would fall too low. They could not eat fast enough to get the energy needed to keep themselves warm.

Challenge 494, page 231: It is about 10^{-9} that of the Earth.

Challenge 496, page 231: The thickness of the folds in the brain, the bubbles in the lung, the density of blood vessels and the size of biological cells.

Challenge 497, page 231: The mercury vapour above the liquid gets saturated.

Challenge 498, page 231: A dedicated NASA project studies this question. Figure 386 gives an example comparison. You can find more details on their website.

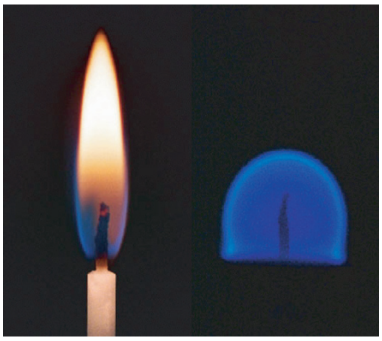


Figure 386 A candle on Earth and in microgravity (nasa)

Challenge 499, page 231: The risks due to storms and the financial risks are too large.

Challenge 500, page 231: The vortex in the tube is cold in the middle and hot at its outside; the air from the middle is sent to one end and the air from the outside to the other. The heating of the outside is due to the work that the air rotating inside has to do on the air outside to get a rotation that eats up angular momentum. For a detailed explanation, see the beautiful text by Mark P. Silverman, *And Yet it Moves: Strange Systems and Subtle Questions in Physics*, Cambridge University Press, 1993, p. 221.

Challenge 501, page 231: Egg white hardens at 70°C, egg white at 65 to 68°C. Cook an egg at the latter temperature, and the feat is possible.

Challenge 505, page 232: This is also true for the shape human bodies, the brain control of human motion, the growth of flowers, the waves of the sea, the formation of clouds, the processes leading to volcano eruptions, etc.

Challenge 510, page 236: There are many more butterflies than tornadoes. In addition, the belief in the butterfly effect completely neglects an aspect of nature that is essential for self-organization: friction and dissipation. The butterfly effect, assumed it did exist, requires that dissipation is neglected. There is no experimental basis for the effect, it has never been observed.

Challenge 520, page 240: All three statements are hogwash. A drag coefficient implies that the cross area of the car is known to the same precision. This is actually extremely difficult to measure and to keep constant. In fact, the value 0.375 for the Ford Escort was a cheat, as many other measurements showed. The fuel consumption is even more ridiculous, as it implies that fuel volumes and distances can be measured to that same precision. Opinion polls are taken by phoning at most 2000 people; due to the difficulties in selecting the right representative sample, that gives a precision of at most 3 %.

Challenge 521, page 241: No. Nature does not allow more than about 20 digits of precision, as we will discover later in our walk. That is not sufficient for a standard book. The question whether such a number can be part of its own book thus disappears.

Challenge 523, page 241: Every measurement is a comparison with a standard; every comparison requires light or some other electromagnetic field. This is also the case for time measurements.

Challenge 524, page 242: Every mass measurement is a comparison with a standard; every comparison requires light or some other electromagnetic field.

Challenge 525, page 242: Angle measurements have the same properties as length or time measurements.

Challenge 526, page 249: A cone or a hyperboloid also look straight from all directions, provided the positioning is correct. One thus needs not only to turn the object, but also to displace it. The best method to check planarity is to use interference between an arriving and a departing coherent beam of light. If the fringes are straight, the surface is planar. (How do you ensure the wavefront of the light beam is planar?)

Challenge 527, page 250: A fraction of infinity is still infinite.

Challenge 528, page 250: The time at which the Moon Io enters the shadow in the second measurement occurs about 1000 s later than predicted from the first measurement. Since the Earth is about $3 \cdot 10^{11}$ m further away from Jupiter and Io, we get the usual value for the speed of light.

Challenge 529, page 251: To compensate for the aberration, the telescope has to be inclined *along* the direction of motion of the Earth; to compensate for parallaxis, *against* the motion.

Challenge 530, page 251: Otherwise the velocity sum would be larger than c.

Challenge 531, page 251: The drawing shows it. Observer, Moon and Sun form a triangle. When the Moon is half full, the angle at the Moon is a right angle. Thus the distance ration can be determined, though not easily, as the angle at the observer is very near a right angle as well.

Challenge 532, page 251: There are Cat's-eyes on the Moon; they are used to reflect laser light pulses sent there through telescopes. The timing of the round trip then gives the distance to the Moon. Of course, absolute distance is not know to high precision, but the variations are. The thickness of the atmosphere is the largest source of error.

Challenge 533, page 252: Fizeau used a mirror about 8.6 km away. As the picture shows, he only had to count the teeth of his cog-wheel and measure its rotation speed when the light goes in one direction through one tooth and comes back to the next.

Challenge 534, page 253: The time must be shorter than $T = l/c$, in other words, shorter than 30 ps; it was a *gas* shutter, not a solid one. It was triggered by a light pulse extracted from the one to be photographed; for certain materials, such as the used gas, strong light can lead to bleaching, so that they become transparent. For more details about the shutter and its neat trigger technique, see the paper by the authors.

Challenge 535, page 253: Just take a photograph of a lightning while moving the camera horizontally. You will see that a lightning is made of several discharges; the whole shows that lightning is much slower than light.

If lightning moved only nearly as fast as light itself, the Doppler effect would change it colour depending on the angle at which we look at it, compared to its direction of motion. A nearby lightning would change colour from top to bottom.

Challenge 536, page 254: The fastest lamps were subatomic particles, such as muons, which decay by emitting a photon, thus a tiny flash of light. However, also some stars emit fasts jets of matter, which move with speeds comparable to that of light.

Challenge 537, page 254: For three hours, this gives a speed difference of 0.5 m/s.

Challenge 540, page 256: The spatial coordinate of the event at which the light is reflected is $c(k^2 - 1)T/2$; the time coordinate is $(k^2 + 1)T/2$. Their ratio must be v. Solving for k gives the result.

Challenge 541, page 257: The motion of radio waves, infrared, ultraviolet and gamma rays is also unstoppable. Another past suspect, the neutrino, has been found to have mass and to be thus in principle stoppable. The motion of gravity is also unstoppable.

Challenge 543, page 259: $\lambda_R/\lambda_S = \gamma$.

Challenge 544, page 259: To change from bright red (650 nm) to green (550 nm), $v = 0.166c$ is necessary.

Challenge 545, page 260: People measure the shift of spectral lines, such as the shift of the so-called Lyman-α line of hydrogen, that is emitted (or absorbed) when a free electron is captured
Page 703
(or ejected) by a proton. It is one of the famous Fraunhofer lines.

Challenge 546, page 260: The speeds are given by

$$v/c = \frac{(z+1)^2 - 1}{(z+1)^2 + 1} \tag{848}$$

which implies $v(z = -0.1) = 31\,\mathrm{Mm/s} = 0.1c$ *towards* the observer and $v(z = 5) = 284\,\mathrm{Mm/s} = 0.95c$ *away* from the observer.

A red-shift of 6 implies a speed of $0.96c$; such speeds appear because, as we will see in the section of general relativity, far away objects recede from us. And high red-shifts are observed only for objects which are extremely far from Earth, and the faster the further they are away. For a red-shift of 6 that is a distance of several thousand million light years.

Challenge 547, page 260: No Doppler effect is seen for a distant observer at rest with respect to the large mass. In other cases there obviously is a Doppler effect, but it is not due to the deflection.

Challenge 548, page 260: Sound speed is not invariant of the speed of observers. As a result, the Doppler effect for sound even confirms – within measurement differences – that time is the *same* for observers moving against each other.

Challenge 551, page 262: Inside colour television tubes (they use higher voltages than black and white ones), electrons are described by $v/c \approx \sqrt{2 \cdot 30/511}$ or $v \approx 0.3c$.

Challenge 552, page 262: If you can imagine this, publish it. Readers will be delighted to hear the story.

Challenge 559, page 265: Redrawing Figure 132 on page 256 for the other observer makes the point.

Challenge 560, page 265: The human value is achieved in particle accelerators; the value in nature is found in cosmic rays of the highest energies.

Challenge 561, page 267: The set of events behaves like a manifold, because it behaves like a four-dimensional space: it has infinitely many points around any given starting point, and distances behave as we are used to, limits behave as we are used to. It differs by one added dimension, and by the sign in the definition of distance; thus, properly speaking, it is a Riemannian manifold.

Challenge 562, page 267: Infinity is obvious, as is openness. Thus the topology equivalence can be shown by imagining that the manifold is made of rubber and wrapped around a sphere.

Challenge 563, page 268: The light cone remains unchanged; thus causal connection as well.

Challenge 564, page 269: In such a case, the division of space-time around an inertial observer into future, past and elsewhere would not hold any more, and the future could influence the past (as seen from another observer).

Challenge 569, page 272: Send a light signal from the first clock to the second clock and back. Take the middle time between the departure and arrival, and then compare it with the time at the reflection. Repeat this a few times. See also Figure 132.

Challenge 573, page 273: Not with present experimental methods.

Challenge 577, page 274: The light cannot stay on at any speed, if the glider is shorter than the gap. This is strange, because the bar does not light the lamp even at high speeds, even though in the frame of the bar there is contact at both ends. The reason is that in this case there is not enough time to send the signal to the battery that contact is made, so that the current cannot start flowing.

Assume that current flows with speed u, which is of the order of c. Then, as Dirk Van de Moortel showed, the lamp will go off if the glider length l_{glider} and the gap length l_{gap} obey $l_{\mathrm{glider}}/l_{\mathrm{gap}} < \gamma(u + v)/u$. See also the cited reference.

Why are the debates often heated? Some people will (falsely) pretend that the problem is unphysical; other will say that Maxwell's equations are needed. Still others will say that the problem is absurd, because for larger lengths of the glider, the on/off answer depends on the precise speed value. However, this actually is the case in this situation.

Challenge 578, page 274: Yes, the rope breaks; in accelerated cars, distance changes, as shown later on in the text.

Challenge 579, page 274: The submarine will sink. The fast submarine will even be heavier, as his kinetic energy adds to his weight. The contraction effect would make it lighter, as the captain says, but by a smaller amount. The total weight – counting upwards as positive – is given by $F = -mg(\gamma - 1/\gamma)$.

Challenge 580, page 274: A relativistic submarine would instantly melt due to friction with the water. If not, it would fly of the planet because it moves faster than the escape velocity. And produce several other disasters.

Challenge 581, page 275: The question confuses observation of Lorentz contraction and its measurement. A relativistic pearl necklace does get shorter, but the shortening can only be measured, not photographed. The measured sizes of the pearls are flattened ellipsoids relativistic speeds. The observed necklace consists of overlapping spheres.

Challenge 584, page 278: Yes, ageing in a valley is slowed compared to mountain tops. However, the proper sensation of time is not changed. The reason for the appearance of grey hair is not known; if the timing is genetic, the proper time at which it happens is the same in either location.

Challenge 585, page 278: There is no way to put an observer at the specified points. Proper velocity can only be defined for observers, i.e., for entities which can carry a clock. That is not the case for images.

Challenge 587, page 280: Most interestingly, the horizon can easily move faster than light, if you move your head appropriately, as can the end of the rainbow.

Challenge 590, page 283: Relativity makes the arguments of challenge 126 watertight.

Challenge 594, page 285: The lower collision in Figure 155 shows the result directly, from energy conservation. For the upper collision the result also follows, if one starts from momentum conservation $\gamma mv = \Gamma MV$ and energy conservation $(gamma + 1)m = \Gamma M$.

Challenge 595, page 286: Annihilation of matter and antimatter.

Challenge 601, page 289: Just turn the left side of Figure 158 a bit in anti-clockwise direction.

Challenge 603, page 291: Probably not, as all relations among physical quantities are known now. However, you might check for yourself; one might never know. It is worth to mention that the maximum force in nature was discovered (in this text) after remaining hidden for over 80 years.

Challenge 604, page 293: Write down the four-vectors U' and U and then extract v' as function of v and the relative coordinate speed V. Then rename the variables.

Challenge 605, page 293: Any motion with light speed.

Challenge 606, page 293: $b^0 = 0, b^i = \gamma^2 a_i$.

Challenge 609, page 294: For ultrarelativistic particles, like for massless particles, one has $E = pc$.

Challenge 610, page 295: Hint: evaluate P_1 and P_2 in the rest frame of one particle.

Challenge 611, page 295: Use the definition $\mathbf{f} = \mathrm{d}\mathbf{p}/\mathrm{d}t$ and the relation $\mathbf{KU} = 0 = \mathbf{fv} - \mathrm{d}E/\mathrm{d}t$ valid for rest-mass preserving forces.

Challenge 634, page 304: The energy contained in the fuel must be comparable to the rest mass of the motorbike, multiplied by c^2. Since fuel contains much more mass than energy, that gives a big problem.

Challenge 642, page 307: Yes, it is true.

Challenge 645, page 308: Yes; however, the effect is minimal and depends on the position of the Sun. In fact, what is white at one height is not white at another.

Challenge 647, page 308: Locally, light moves with speed c.

Challenge 648, page 309: Away from Earth, g decreases; it is effectively zero over most of the distance.

Challenge 649, page 309: Light is necessary to determine distance *and* to synchronize clocks; thus there is no way to measure the speed of light from one point to another alone. The reverse motion needs to be included. However, some statements on the one-way speed of light can still be made (see http://math.ucr.edu/home/baez/physics/Relativity/SR/experiments.html). All experiments on the one-way speed of light performed so far are consistent with an isotropic value that is equal to the two-way velocity. However, no experiment is able to rule out a group of theories in which the one-way speed of light is anisotropic and thus different from the two-way speed. All theories from this group have the property that the *round-trip* speed of light is isotropic in any inertial frame, but the *one-way* speed is isotropic only in a preferred 'ether' frame. In all of these theories, in all inertial frames, the effects of slow clock transport exactly compensate the effects of the anisotropic one-way speed of light. All these theories are experimentally indistinguishable from special relativity. In practice, therefore, the one-way speed of light has been measured and is constant. But a small option remains.

Challenge 650, page 310: See the cited reference. The factor 2 was forgotten there; can you deduce it?

Challenge 653, page 311: Though there are many publications pretending to study the issue, there are also enough physicists who notice the impossibility. Measuring a variation of the speed of light is not much far from measuring the one way speed of light: it is not possible. However, the debates on the topic are heated; the issue will take long to be put to rest.

Challenge 654, page 319: The inverse square law of gravity does not comply with the maximum speed principle; it is not clear how it changes when one changes to a moving observer.

Challenge 655, page 323: Take a surface moving with the speed of light, or a surface defined with a precision smaller than the Planck length.

Challenge 660, page 335: For example, it is possible to imagine a surface that has such an intricate shape that it will pass all atoms of the universe at almost the speed of light. Such a surface is not physical, as it is impossible to imagine observers on all its points that move in that way all at the same time.

Challenge 667, page 345: They are accelerated upwards.

Challenge 668, page 345: In everyday life, (a) the surface of the Earth can be taken to be flat, (b) the vertical curvature effects are negligible, and (c) the lateral length effects are negligible.

Challenge 672, page 346: For a powerful bus, the acceleration is $2\,\mathrm{m/s^2}$; in 100 m of acceleration, this makes a relative frequency change of $2.2 \cdot 10^{-15}$.

Challenge 674, page 346: Yes; any absorption of light by a body or any emission of light by a lamp are an example.

Challenge 677, page 346: For a beam of light, in both cases the situation is described by an environment in which masses 'fall' against the direction of motion. If the Earth and the train walls were not visible – for example if they were hidden by mist – there would not be any way to determine by experiment which situation is which. Or again, if an observer would be enclosed in a box, he could not distinguish between constant acceleration or constant gravity. (Important: this impossibility only applies if the observer has negligible size!)

Challenge 683, page 348: Both fall towards the centre of the Earth. Orbiting particles are also in free fall; their relative distance changes as well, as explained in the text.

Challenge 686, page 350: Such a graph would need four or even 5 dimensions.

Challenge 688, page 351: The energy due to the rotation can be neglected compared with all other energies in the problem.

Challenge 696, page 356: Different nucleons, different nuclei, different atoms and different molecules have different percentages of binding energies relative to the total mass.

Challenge 699, page 358: Let the device fall. The elastic rubber then is strong enough to pull the ball into the cup. See M.T. Westra, Einsteins verjaardagscadeau, *the nederlands tijdschrift voor natuurkunde* **69**, p. 109, April 2003.

Challenge 705, page 358: They use a spring scale, and measure the oscillation time. From it they deduce their mass.

Challenge 706, page 358: After about half an hour.

Challenge 710, page 359: With $\hbar$ as smallest angular momentum one get about 100 Tm.

Challenge 711, page 359: No. The diffraction of the beams does not allow it. Also quantum theory makes this impossible; bound states of massless particles, such as photons, are not stable.

Challenge 713, page 360: The orbital radius is 4.2 Earth radii; that makes *c.* 38 μs every day.

Challenge 714, page 360: To be honest, the experiments are not consistent. They assume that some other property of nature is constant – such as atomic size – which in fact also depends on G. More on this issue on page 463.

Challenge 715, page 361: Of course other spatial dimensions could exist which can be detected only with the help of measurement apparatuses. For example, hidden dimensions could appear at energies not accessible in everyday life.

Challenge 725, page 366: Since there is no negative mass, gravitoelectric fields cannot be neutralized. In contrast, electric fields can be neutralized around a metallic conductor with a Faraday cage.

Challenge 738, page 374: One needs to measure the timing of pulses which cross the Earth at different gravitational wave detectors on Earth.

Challenge 754, page 380: No; a line cannot have intrinsic curvature. A torus is indeed intrinsically curved; it cannot be cut open to a flat sheet of paper.

Challenge 792, page 399: Indeed, in general relativity gravitational energy cannot be localized in space, in contrast to what one expects and requires from an interaction.

Challenge 808, page 418: The rabbit observes that all other rabbits seem to move away from him.

Challenge 878, page 456: Any device that uses mirrors requires electrodynamics; without electrodynamics, mirrors are impossible.

Challenge 880, page 458: The hollow Earth theory is correct if usual distance are consistently changed to $r_{\mathrm{he}} = R^2_{\mathrm{Earth}}/r$. This implies a quantum of action that decreases towards the centre of the hollow sphere. Then there is no way to prefer one description over the other, except for reasons of simplicity.

Challenge 865, page 450: This happens in the same way that the static electric field comes out of a charge. In both cases, the transverse fields do not get out, but the longitudinal fields do. Quantum theory provides the deeper reason. Real radiation particles, which are responsible for free, transverse fields, cannot leave a black hole because of the escape velocity. However, virtual particles can, as their speed is not bound by the speed of light. All static, longitudinal fields are produced by virtual particles. In addition, there is a second reason. Classical field can come out of a black hole because for an outside observer everything making it up is continuously falling, and nothing has actually crossed the horizon. The field sources thus are not yet out of reach.

Challenge 869, page 451: The description says it all. A visual impression can be found in the room on black holes in the 'Deutsches Museum' in München.

Challenge 887, page 482: The liquid drops have to detach from the flow exactly inside the metal counter-electrodes. Opel simply earthed the metal piece they had built into the cars without any contact to the rest of the car.

Challenge 888, page 483: A lot of noise while banging up and down.

Challenge 891, page 485: The field at a distance of 1 m from an electron is 1.4 nV/m.

Challenge 892, page 486: A simple geometrical effect: anything flowing out homogeneously from a sphere diminishes with the square of the distance.

Challenge 893, page 486: One has $F = \alpha\hbar c N_A^2/4R^2 = 3 \cdot 10^{12}$ N, an enormous force, corresponding to the weight of 300 million tons. It shows the enormous forces that keep matter together. Obviously, there is no way to keep 1 g of positive charge together, as the repulsive forces among the charges would be even larger.

Challenge 895, page 487: No; they only separate charges and pump them around.

Challenge 896, page 487: Uncharged bodies can attract each other if they are made of charged constituents neutralizing each other, and if the charges are constrained in their mobility. The charge fluctuations then lead to attraction. Most molecules interact among each other in this way; such forces are also at the basis of surface tension in liquids and thus of droplet formation.

Challenge 898, page 488: The ratio q/m of electrons and that of the free charges inside metals is not exactly the same.

Challenge 1003, page 545: See challenge 551.

Challenge 1004, page 546: The electrons move slowly, but the speed of electrical signals is given by the time at which the electrons move. Imagine long queue of cars (representing electrons) waiting in front of a red traffic light. All drivers look at the light. As soon as it turns green, everybody starts driving. Even though the driving speed might be only 10 m/s, the speed of traffic flow onset was that of light. It is this latter speed which is the speed of electrical signals.

Water pipes tell the same story. A long hose provides water almost in the same instant as the tap is opened, even if the water takes a long time to arrive from the tap to the end of the hose. The speed with which the water reacts is gives by the speed for pressure waves in water. Also for water hoses the signal speed, roughly given by the sound speed in water, is much higher than the speed of the water flow.

Challenge 906, page 496: The dual field $*F$ is defined on page 504.

Challenge 907, page 496: Scalar products of four vectors are always, by construction, Lorentz invariant quantities.

Challenge 913, page 498: Usually, the cables of high voltage lines are too warm to be comfortable.

Challenge 914, page 498: Move them to form a T shape.

Challenge 915, page 498: For four and more switches, on uses inverters; an inverter is a switch with two inputs and two outputs which in one position, connects first and second input to first and second output respectively, and in the other position connects the first input to the second output and vice versa. (There are other possibilities, though; wires can be saved using electromagnetic relay switches.) For three switches, there is a simpler solution than with inverters.

Challenge 917, page 499: It is possible; however, the systems so far are not small and are dangerous for human health. The idea to collect solar power in deep space and then beam it to the Earth as microwaves has often been aired. Finances and dangers have blocked it so far.

Challenge 918, page 499: Glue two mirrors together at a right angle. Or watch yourself on TV using a video camera.

Challenge 952, page 513: The image flips up: a 90 degree rotation turns the image by 180 degrees.

Challenge 919, page 499: This is again an example of combined triboluminescence and triboelectricity. See also the websites http://scienceworld.wolfram.com/physics/Triboluminescence.html and http://www.geocities.com/RainForest/9911/tribo.htm.

Challenge 921, page 500: Pepper is lighter than salt, and thus reacts to the spoon before the salt does.

Challenge 922, page 501: For a wavelength of 546.1 nm (standard green), that is a bit over 18 wavelengths.

Challenge 923, page 501: The angular size of the Sun is too large; diffraction plays no role here.

Challenge 924, page 501: Just use a high speed camera.

Challenge 1005, page 546: One can measure current fluctuations, or measure smallest charges, showing that they are always multiples of the same unit. The latter method was used by Millikan.

Challenge 925, page 501: The current flows perpendicularly to the magnetic field and is thus deflected. It pulls the whole magnet with it.

Challenge 926, page 502: Light makes seven turns of the Earth in one second.

Challenge 1008, page 546: Earth's potential would be $U = -q/(4\pi\varepsilon_o R) = 60\,\mathrm{MV}$, where the number of electrons in water must be taken into account.

Challenge 929, page 502: The most simple equivalent to a coil is a rotating mass being put into rotation by the flowing water. A transformer would then be made of two such masses connected through their axis.

Challenge 932, page 503: The charged layer has the effect that almost only ions of one charge pass the channels. As a result, charges are separated on the two sides of the liquid, and a current is generated.

Challenge 936, page 506: Some momentum is carried away by the electromagnetic field.

Challenge 937, page 506: Field lines and equipotential surfaces are always orthogonal to each other. Thus a field line cannot cross an equipotential surface twice.

Challenge 948, page 512: Just draw a current through a coil with its magnetic field, then draw the mirror image of the current and redraw the magnetic field.

Challenge 949, page 512: Other asymmetries in nature include the helicity of the DNA molecules making up the chromosomes and many other molecules in living systems, the right hand preference of most humans, the asymmetry of fish species which usually stay flat on the bottom of the seas.

Challenge 950, page 513: This is not possible at all using gravitational or electromagnetic systems or effects. The only way is to use the weak nuclear interaction, as shown in the chapter on the nucleus.

Challenge 951, page 513: The Lagrangian does not change if one of the three coordinates is changed by its negative value.

Challenge 953, page 514: Imagine E and B as the unite vectors of two axes in complex space. Then any rotation of these axes is also a generalized duality symmetry.

Challenge 957, page 518: In every case of interference, the energy is redistributed into other directions. This is the general rule; sometimes it is quite tricky to discover this other direction.

Challenge 958, page 518: The author regularly sees about 7 lines; assuming that the distance is around $20\,\mu\mathrm{m}$, this makes about $3\,\mu\mathrm{m}$ per line. The wavelength must be smaller than this value and the frequency thus larger than 100 THz. The actual values for various colours are given in the table of the electromagnetic spectrum.

Challenge 960, page 519: He noted that when a prism produces a rainbow, a thermometer placed in the region after the colour red shows a temperature rise.

Challenge 961, page 520: Light reflected form a water surface is partly polarized. Mirages are not.

Challenge 963, page 520: Drawing them properly requires four dimensions; and there is no analogy with two-dimensional waves. It is not easy to picture them. The direction of oscillation of the fields rotates as the wave advances. The oscillation direction thus forms a spiral. Picturing the rest of the wave is not impossible, but not easy.

Challenge 966, page 524: Such an observer would experience a wavy but static field, which cannot exist, as the equations for the electromagnetic field show.

Challenge 967, page 525: Syrup shows an even more beautiful effect in the following setting. Take a long transparent tube closed at one end and fill it with syrup. Shine a red helium–neon laser into the tube from the bottom. Then introduce a linear polarizer into the beam: the light seen in the tube will form a spiral. By rotating the polarizer you can make the spiral advance or retract. This effect, called the *optical activity* of sugar, is due to the ability of sugar to rotate light polarization and to a special property of plants: they make only one of the two mirror forms of sugar.

Challenge 969, page 526: The 1 mm beam would return 1000 times as wide as the 1 m beam. A perfect 1 m-wide beam of green light would be 209 m wide on the Moon; can you deduce this result from the (important) formula that involves distance, wavelength, initial diameter and final diameter? Try to guess this beautiful formula first, and then deduce it. In reality, the values are a few times larger than the theoretical minimum thus calculated. See the http://www.csr.utexas.edu/mlrs and http://ilrs.gsfc.nasa.gov websites.

Challenge 970, page 526: The answer should lie between one or two dozen kilometres, assuming ideal atmospheric circumstances.

Challenge 975, page 529: A surface of 1 m^2 perpendicular to the light receives about 1 kW of radiation. It generates the same pressure as the weight of about 0.3 mg of matter. That generates 3 μPa for black surfaces, and the double for mirrors.

Challenge 977, page 529: The shine side gets twice the momentum transfer as the black side, and thus should be pushed backwards.

Challenge 980, page 530: A polarizer can do this.

Challenge 983, page 531: The interference patterns change when colours are changed. Rainbows also appear because different colours are due to different frequencies.

Challenge 985, page 531: The full rainbow is round like a circle. You can produce one with a garden hose, if you keep the hose in your hand while you stand on a chair, with your back to the evening Sun. (Well, one small part is missing; can you imagine which part?) The circle is due to the spherical shape of droplets. If the droplets were of different shape, *and* if they were all aligned, the rainbow would have a different shape than a simple circle.

Challenge 989, page 533: Film a distant supernova explosion and check whether it happens at the same time for each colour separately.

Challenge 991, page 535: The first part of the forerunner is a feature with the shortest possible effective wavelength; thus it is given by taking the limit for infinite frequency.

Challenge 992, page 535: The light is pulsed; thus it is the energy velocity.

Challenge 993, page 535: Inside matter, the energy is transferred to atoms, then back to light, then to the next atoms, etc. That takes time and slows down the propagation.

Challenge 996, page 538: This is true even in general relativity, when the bending of the vacuum is studied.

Challenge 1010, page 547: Almost no light passes; the intensity of the little light that is transmitted depends exponentially on the ratio between wavelength and hole diameter. One also says that after the hole there is an evanescent wave.

Challenge 1012, page 547: The angular momentum was put into the system when it was formed. If we bring a point charge from infinity along a straight line to its final position close to a magnetic dipole, the magnetic force acting on the charge is not directed along the line of motion. It therefore creates a non-vanishing torque about the origin. See J.M. Aguirregabiria & A. Hernandez, The Feynman paradox revisited, *European Journal of Physics* 2, pp. 168–170, 1981.

Challenge 1014, page 547: Leakage currents change the picture. The long term voltage ratio is given by the leakage resistance ratio $V_1/V_2 = R_1/R_2$, as can be easily verified in experiments.

Challenge 1015, page 547: There is always a measurement error when measuring field values, even when measuring a 'vanishing' electromagnetic field.

Challenge 1016, page 548: The green surface seen at a low high angle is larger than when seen vertically, where the soil is also seen; the soil is covered by the green grass in low angle observation.

Challenge 1017, page 548: The charges in a metal rearrange in a way that the field inside remains vanishing. This makes cars and aeroplanes safe against lightning. Of course, if the outside field varies so quickly that the rearrangement cannot follow, fields *can* enter the Faraday cage. (By the way, also fields with long wavelengths penetrate metals; remote controls regularly use frequencies of 25 kHz to achieve this.) However, one should wait a bit before stepping out of a car after lightning has hit, as the car is on rubber wheels with low conduction; waiting gives the charge time to flow into the ground.

For gravity and solid cages, mass rearrangement is not possible, so that there is no gravity shield.

Challenge 1023, page 549: Of course not, as the group velocity is not limited by special relativity. The energy velocity is limited, but is not changed in this experiments.

Challenge 968, page 526: A light microscope is basically made of two converging lenses. One lens – or lens system – produces an enlarged real image and the second one produces an enlarged virtual image of the previous real image. Figure 387 also shows that microscopes always turn images upside down. Due to the wavelength of light, light microscopes have a maximum resolution of about 1 μm. Note that the magnification of microscopes is unlimited; what is limited is their resolution. This is exactly the same behaviour shown by digital images. The *resolution* is simply the size of the smallest possible pixel that makes sense.

Challenge 1025, page 549: The Prussian explorer Alexander von Humboldt extensively checked this myth in the nineteenth century. He visited many mine pits and asked countless mine workers in Mexico, Peru and Siberia about their experiences. He also asked numerous chimney-sweeps. Neither him nor anybody else had ever seen the stars during the day.

Challenge 1026, page 549: The number of photons times the quantum of action $\hbar$.

Challenge 1029, page 550: The charging stops because a negatively charged satellite repels electrons and thus stops any electron collecting mechanism. Electrons are captured more frequently than ions because it is easier for them than for ions to have an inelastic collision with the satellite, due to their larger speed at a given temperature.

Challenge 1030, page 550: Any loss mechanism will explain the loss of energy, such as electrical resistance or electromagnetic radiation. After a fraction of a second, the energy will be lost. This

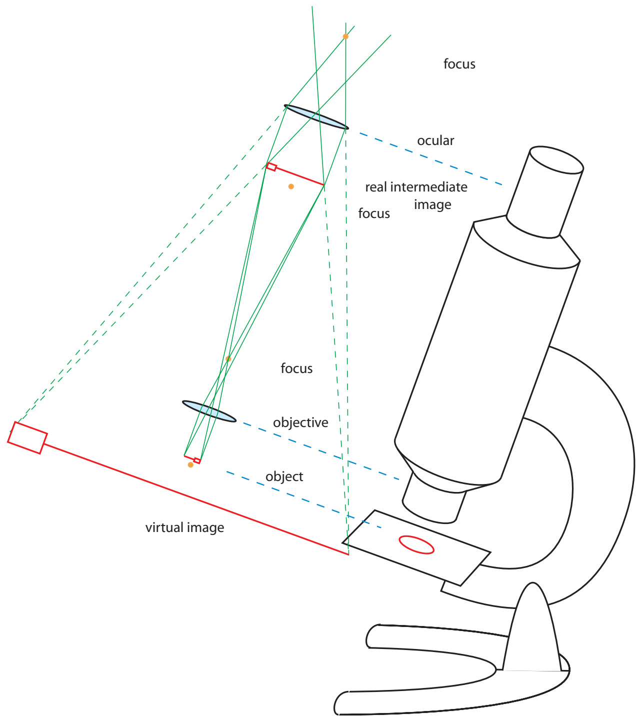


Figure 387 Two converging lenses make a microscope

little problem is often discussed on the internet.

Challenge 1032, page 550: Show that even though the radial magnetic field of a spherical wave is vanishing by definition, Maxwell's equations would require it to be different from zero. Since electromagnetic waves are transversal, it is also sufficient to show that it is impossible to comb a hairy sphere without having a (double) vortex or two simple vortices. Despite these statements, quantum theory changes the picture somewhat: the emission probability of a photon from an excited atom in a degenerate state is spherically symmetric exactly.

Challenge 1033, page 552: The human body is slightly conducting and changes the shape of the field and thus effectively short circuits it. Usually, the field cannot be used to generate energy, as the currents involved are much too small. (Lightning bolts are a different story, of course. They are due – very indirectly – to the field of the Earth, but they are too irregular to be used consistently. Franklin's lightning rod is such an example.)

Challenge 997, page 539: Not really; a Cat's-eye uses two reflections at the sides of a cube. A living cat's eye has a large number of reflections. The end effect is the same though: light returns back to the direction it came from.

Challenge 998, page 539: There is a blind spot in the eye; that is a region in which images are not perceived. The brain than assumes that the image at that place is the same than at its borders. If a spot falls exactly inside it, it disappears.

Challenge 1000, page 542: The eye and vision system subtract patterns that are constant in time.

Challenge 1002, page 543: A hologram is always transparent; one can always see the background through the hologram. A hologram thus always gives an impression similar to what movies usually show as ghosts.

Challenge 1037, page 555: This should be possible in the near future; but both the experiment, which will probably measure brain magnetic field details, and the precise check of its seriousness will not be simple.

Challenge 1044, page 558: Any new one is worth a publication.

Challenge 1045, page 562: Sound energy is also possible, as is mechanical work.

Challenge 1046, page 563: Space-time deformation is not related to electricity; at least at every-day energies. Near Planck energies, this might be different, but nothing has been predicted yet.

Challenge 1048, page 565: Ideal absorption is blackness (though it can be redness or whiteness at higher temperatures).

Challenge 1049, page 565: Indeed, the Sun emits about $4 \cdot 10^{26}$ W from its mass of $2 \cdot 10^{30}$ kg, about 0.2 mW/kg. The adult human body (at rest) emits about 100 W (you can check this in bed at night), thus about 1.2 W/kg per ton. This is about 6000 times more than the Sun.

Challenge 1050, page 565: The average temperature of the Earth is thus 287 K. The energy from the Sun is proportional to the fourth power of the temperature. The energy is spread (roughly) over half the Earth's surface. The same energy, at the Sun's surface, comes from a much smaller surface, given by the same angle as the Earth subtends there. We thus have $E \sim 2\pi R^2_{\mathrm{Earth}} T^4_{\mathrm{Earth}} = T^4_{\mathrm{Sun}} R^2_{\mathrm{Earth}} \alpha^2$, where α is half the angle subtended by the Sun. As a result, the temperature of the Sun is estimated to be $T_{\mathrm{Sun}} = (T^4_{\mathrm{Earth}}/\alpha^2)^{0.25} = 4\,\mathrm{kK}$.

Challenge 1057, page 567: At high temperature, all bodies approach black bodies. The colour is more important than other colour effects. The oven and the objects have the same temperature. Thus they cannot be distinguished from each other. To do so nevertheless, illuminate the scene with powerful light and then take a picture with small sensitivity. Thus one always needs bright light to take pictures of what happens inside fires.

Challenge 1062, page 586: The issue is: is the 'universe' a concept? More about this issue in the third part of the text.

Challenge 1064, page 589: When thinking, physical energy, momentum and angular momentum are conserved, and thermodynamic entropy is not destroyed. Any experiment that this would not be so would point to unknown processes. However, there is no evidence for this.

Challenge 1065, page 589: The best method cannot be much shorter than what is needed to describe 1 in 6000 million, or about 34 bits. The Dutch and UK post code systems (including the letters NL or UK) are not far from this value and thus can claim to be very efficient.

Challenge 1066, page 589: For complex systems, when the unknowns are numerous, the advance is thus simply given by the increase in answers. For the universe as a whole, the number of open issues is quite low, as shown on page 887; here there has not been much advance in the last years. But the advance is clearly measurable in this case as well.

Challenge 1067, page 590: Is it possible to use the term 'complete' when describing nature?

Challenge 1070, page 591: There are many baths in series: thermal baths in each light-sensitive cell of the eyes, thermal baths inside the nerves towards the brain and thermal baths inside brain cells.

Challenge 1072, page 591: Yes.

Challenge 1075, page 597: Physicists claim that the properties of objects, of space-time and of interactions form the smallest list possible. However, this list is longer than the one found by linguists! The reason is that physicists have found primitives that do not appear in everyday life. In a sense, the aim of physicists is limited by list of unexplained questions of nature, given on page 887.

Challenge 1076, page 598: Neither has a defined content, clearly stated limits or a domain of application.

Challenge 1077, page 598: Impossible! That would not be a concept, as it has no content. The solution to the issue must be and will be different.

Challenge 1078, page 600: To neither. This paradox shows that such a 'set of all sets' does not exist.

Challenge 1079, page 600: The most famous is the class of all sets that do not contain themselves. This is not a set, but a class.

Challenge 1080, page 601: Dividing cakes is difficult. A just method (in finite many steps) for 3 people, using nine steps, was published in 1944 by Steinhaus, and a fully satisfactory method in the 1960s by Conway. A fully satisfactory method for four persons was found only in 1995; it has 20 steps.

Challenge 1081, page 601: $(x, y) := \{x, \{x, y\}\}$.

Challenge 1082, page 602: Hint: show that any countable list of reals misses at least one number. This was proven for the first time by Cantor. His way was to write the list in decimal expansion and then find a number that is surely not in the list. Second hint: his world-famous trick is called the diagonal argument.

Challenge 1083, page 602: Hint: all reals are limits of series of rationals.

Challenge 1085, page 603: Yes.

Challenge 1086, page 604: There are infinitely many of them. But the smallest is already quite large.

Challenge 1087, page 604: $0 := \varnothing \quad , \quad 1 := \{\varnothing\} \quad , \quad 2 := \{\{\varnothing\}\}$ etc.

Challenge 1088, page 608: Subtraction is easy. Addition is not commutative only for cases when infinite numbers are involved: $\omega + 2 \neq 2 + \omega$.

Challenge 1089, page 608: Examples are $1 - \varepsilon$ or $1 - 4\varepsilon^2 - 3\varepsilon^3$.

Challenge 1090, page 608: The answer is 57; the cited reference gives the details.

Challenge 1091, page 610: $2^{2^{22}}$ and $4^{4^{4^4}}$.

Challenge 1093, page 610: This is not an easy question. The first nontrivial numbers are 7, 23, 47, 59, 167 and 179. See Robert Matthews, Maximally periodic reciprocals, *Bulletin of the Institute of Mathematics and its Applications* 28, pp. 147–148, 1992. Matthews shows that a number n for which $1/n$ generates the maximum of $n - 1$ decimal digits in the decimal expansion is a special sort of prime number that can be deduced from the so-called *Sophie Germain primes* S; one must have $n = 2S + 1$, where both S and $2S + 1$ must be prime and where S mod 20 must be 3, 9, or 11.

Thus the first numbers n are 7, 23, 47, 59, 167 and 179, corresponding to values for S of 3, 11, 23, 29, 83 and 89. In 1992, the largest known S that meets the criteria was

$$S = (39051 \cdot 2^{6002}) - 1 , \qquad (849)$$

a 1812-digit long Sophie Germain prime number that is 3 mod 20. It was discovered by Wilfred Keller. This Sophie Germain prime leads to a prime n with a decimal expansion that is around 10^{1812} digits long before it starts repeating itself. Read your favourite book on number theory to find out more. Interestingly, the solution to this challenge is also connected to that of challenge 1086. Can you find out more?

Challenge 1094, page 610: Klein did not belong to either group. As a result, some of his nastier students concluded that he was not a mathematician at all.

Challenge 1095, page 611: A barber cannot belong to either group; the definition of the barber is thus contradictory and has to be rejected.

Challenge 1096, page 611: See the http://members.shaw.ca/hdhcubes/cube_basics.htm web page for more information on magic cubes.

Challenge 1097, page 611: Such an expression is derived with the intermediate result $(1-2^2)^{-1}$. The handling of divergent series seems absurd, but mathematicians know how to give the expression a defined content. (See Godfrey H. Hardy, *Divergent Series*, Oxford University Press, 1949.) Physicists often use similar expressions without thinking about them, in quantum field theory.

Challenge 1098, page 620: 'All Cretans lie' is *false*, since the opposite, namely 'some Cretans say the truth' is true in the case given. The trap is that the opposite of the original sentence is usually, but *falsely*, assumed to be 'all Cretans say the truth'.

Challenge 1099, page 620: The statement cannot be false, due to the first half and the 'or' construction. Since it is true, the second half must be true and you are an angel.

Challenge 1107, page 621: The light bulb story seems to be correct. The bulb is very weak, so that the wire is not evaporating.

Challenge 1108, page 626: Only induction allows to make use of similarities and thus to define concepts.

Page 966 **Challenge 1110**, page 628: Yes, as we shall find out.

Challenge 1111, page 629: Yes, as observation implies interaction.

Challenge 1112, page 629: Lack of internal contradictions means that a concept is valid as a thinking tool; as we use our thoughts to describe nature, mathematical existence is a specialized version of physical existence, as thinking is itself a natural process. Indeed, mathematical concepts are also useful for the description of the working of computers and the like.

Another way to make the point is to stress that all mathematical concepts are built from sets and relations, or some suitable generalizations of them. These basic building blocks are taken from our physical environment. Sometimes the idea is expressed differently; many mathematicians have acknowledged that certain mathematical concepts, such as natural numbers, are taken directly from experience.

Challenge 1113, page 629: Examples are Achilles, Odysseus, Mickey Mouse, the gods of polytheism and spirits.

Challenge 1115, page 631: Torricelli made vacuum in a U-shaped glass tube, using mercury, the same liquid metal used in thermometers. Can you imagine how? A more difficult question: where did he get mercury from?

Challenge 1116, page 632: Stating that something is infinite can be allowed, if the statement is falsifiable. An example is the statement 'There are infinitely many mosquitoes.'

Other statements are not falsifiable, such as 'The universe continue without limit behind the horizon.' Such a statement is a belief, not a fact.

Challenge 1117, page 634: They are not sets either and thus not collections of points.

Challenge 1118, page 635: There is still no possibility to interact with all matter and energy, as this includes oneself.

Challenge 1119, page 640: No. There is only a generalization encompassing the two.

Challenge 1120, page 641: An explanation of the universe is not possible, as the term explanation require the possibility to talk about systems outside the one under consideration. The universe is not part of a larger set.

Challenge 1121, page 641: Both can in fact be seen as two sides of the same argument: there is no other choice;there is only one possibility. The rest of nature shows that it has to be that way, as everything depends on everything.

Challenge 1122, page 656: Classical physics fails in explaining any material property, such as colour or softness. Material properties result from nature's interactions; they are inevitably quantum. Explanations always require particles and their quantum properties.

Challenge 1123, page 657: Classical physics allows any observable to change *smoothly* with time. There is no minimum value for any observable physical quantity.

Challenge 1125, page 658: The simplest length is $\sqrt{2G\hbar/c^3}$. The factor 2 is obviously not fixed; it is explained later on. Including it, this length is the smallest length measurable in nature.

Challenge 1126, page 658: The electron charge is special to the electromagnetic interactions; it does not take into account the nuclear interactions. It is also unclear why the length should be of importance for neutral systems or for the vacuum. On the other hand, it turns out that the differences are not too fundamental, as the electron charge is related to he quantum of action by $e = \sqrt{4\pi\varepsilon_0 \alpha c\hbar}$.

Challenge 1127, page 658: On purely dimensional grounds, the radius of an atom must be

$$r \approx \frac{\hbar^2 4\pi\varepsilon_0}{me^2} \qquad (850)$$

This is about 160 nm; indeed, this guessed equation is simply π times the Bohr radius.

Challenge 1128, page 659: Due to the quantum of action, atoms in all people, be they giants or dwarfs, have the same size. That giants do not exist was shown already by Galilei. The argument is based on the given strength of materials, which thus implies that atoms are the same everywhere. That dwarfs cannot exist is due to the same reason; nature is not able to make people smaller than usual (except in the womb) as this would require smaller atoms.

Page 203

Challenge 1139, page 665: Also photons are indistinguishable. See page 682.

Challenge 1142, page 667: The total angular momentum counts, including the orbital angular momentum. The orbital angular momentum $\mathbf{L}$ is given, using the radius and the linear momentum, $\mathbf{L} = \mathbf{r} \times \mathbf{p}$.

Challenge 1143, page 667: Yes, we could have!

Challenge 1163, page 685: The quantum of action implies that two subsequent observations always differ. Thus the surface of a liquid cannot be at rest.

Challenge 1174, page 701: Use $\Delta E < E$ and $a\ \Delta t < c$.

Challenge 1178, page 702: The difficulties to see hydrogen atoms are due to their small size and their small number of electrons. As a result, hydrogen atoms produce only weak contrasts in X-ray images. For the same reasons it is difficult to image them using electrons; the Bohr radius of hydrogen is only slightly larger than the electron Compton wavelength.

Challenge 1182, page 702: $r = 86$ pm, thus $T = 12$ eV. That compares to the actual value of 13.6 eV. The trick for the derivation of the formula is to use $<\psi|r_x^2|\psi>= \frac{1}{3} <\psi|\mathbf{rr}|\psi>$, a relation valid for states with no orbital angular momentum. It is valid for all coordinates and also for the three momentum observables, as long as the system is non-relativistic.

Challenge 1183, page 702: The fields are crated by neutrons or protons, which have a smaller Compton wavelength.

Challenge 1197, page 713: A change of physical units such that $\hbar = c = e = 1$ would change the value of ε_0 in such a way that $4\pi\varepsilon_o = 1/\alpha = 137.036...$

Challenge 1198, page 719: Point particles cannot be marked; nearby point particles cannot be distinguished, due to the quantum of action.

Challenge 1204, page 722: For a large number of particles, the interaction energy will introduce errors. For very large numbers, the gravitational binding energy will do so as well.

Challenge 1206, page 723: Two write two particles on paper, one has to distinguish them, even if the distinction is arbitrary.

Challenge 1211, page 727: In the way their intestines are folded, in the lines of their hands and other skin lines; often features like black points on the skin are mirror inverted on the two twins.

Challenge 1215, page 733: Angels can be distinguished by name, can talk and can sing; thus they are made of a large number of fermions. In fact, many angels are human sized, so that they do not even fit on the tip of a pin.

Challenge 1223, page 736: Ghosts, like angels, can be distinguished by name, can talk and can be seen; thus they contain fermions. However, they can pass through walls and they are transparent; thus they cannot be made of fermions, but must be images, made of bosons. That is a contradiction.

Challenge 1225, page 742: The loss of non-diagonal elements leads to an increase in the diagonal elements, and thus of entropy.

Challenge 1228, page 747: The energy speed is given by the advancement of the outer two tails; that speed is never larger than the speed of light.

Challenge 1231, page 750: Such a computer requires clear phase relations between components; such phase relations are extremely sensitive to outside disturbances. At present, they do not hold longer than a microsecond, whereas long computer programs require minutes and hours to run.

Challenge 1235, page 756: Any other bath also does the trick, such as the atmosphere, sound vibrations, electromagnetic fields, etc.

Challenge 1236, page 757: The Moon is in contact with baths like the solar wind, falling meteorites, the electromagnetic background radiation of the deep universe, the neutrino flux from the Sun, cosmic radiation, etc.

Challenge 1237, page 758: Spatially periodic potentials have the property. Decoherence then leads to momentum diagonalisation.

Challenge 1238, page 761: A virus is an example. It has no own metabolism. (By the way, the ability of some viruses to form crystals is not a proof that they are not living beings, in contrast to what is often said.)

Challenge 1239, page 762: The navigation systems used by flies are an example.

Challenge 1240, page 763: The thermal energy kT is about 4 zJ and a typical relaxation time is 0.1 ps.

Challenge 1241, page 766: This is not possible at present. If you know a way, publish it. It would help a sad single mother who has to live without financial help from the father, despite a lawsuit, as it was yet impossible to decide which of the two candidates is the right one.

Challenge 1242, page 766: Also identical twins count as different persons and have different fates. Imprinting in the womb is different, so that their temperament will be different. The birth experience will be different; this is the most intense experience of every human, strongly determining his fears and thus his character. A person with an old father is also quite different from that with a young father. If the womb is not that of his biological mother, a further distinction of the earliest and most intensive experiences is given.

Challenge 1243, page 767: Life's chemicals are synthesized inside the body; the asymmetry has been inherited along the generations. The common asymmetry thus shows that all life has a common origin.

Challenge 1244, page 767: Well, men are more similar to chimpanzees than to women. More seriously, the above data, even though often quoted, are wrong. Newer measurements by Roy Britten in 2002 have shown that the difference in genome between humans and chimpanzees is about 5 % (See R. J. Britten, Divergence between samples of chimpanzee and human DNA sequences is 5 %, counting indels, *Proceedings of the National Academy of Sciences* **99**, pp. 13633–13635, 15th of October, 2002.) In addition, though the difference between man and woman is smaller than one whole chromosome, the large size of the X chromosome, compared with the small size of the Y chromosome, implies that men have about 3 % less genetic material than women. However, all men have an X chromosome as well. That explains that still other measurements suggest that all humans share a pool of at least 99.9 % of common genes.

Challenge 1263, page 780: All detectors of light can be called relativistic, as light moves with maximal speed. Touch sensors are not relativistic following the usual sense of the word, as the speeds involved are too small. The energies are small compared to the rest energies; this is the case even if the signal energies are attributed to electrons only.

Challenge 1246, page 767: Since all the atoms we are made of originate from outer space, the answer is yes. But if one means that biological cells came to Earth from space, the answer is no, as cells do not like vacuum. The same is true for DNA.

Challenge 1245, page 767: The first steps are not known yet.

Challenge 1248, page 768: Chemical processes, including diffusion and reaction rates, are strongly temperature dependent. They affect the speed of motion of the individual and thus its chance of survival. Keeping temperature in the correct range is thus important for evolved life forms.

Challenge 1249, page 770: Haven't you tried yet? Physics is an experimental science.

Challenge 1251, page 771: Radioactive dating methods can be said to be based on the nuclear interactions, even though the detection is again electromagnetic.

Challenge 1267, page 784: With a combination of the methods of Table 60 it is possible; but whether there will ever be an organization willing to pay for this to happen is another question.

Challenge 1269, page 788: For example, a heavy mountain will push down the Earth's crust into the mantle, makes it melt on the bottom side, and thus lowers the position of the top.

Challenge 1270, page 788: These developments are just starting; the results are still far from the original one is trying to copy, as they have to fulfil a second condition, in addition to being a 'copy' of original feathers or of latex: the copy has to be cheaper than the original. That is often a much tougher request than the first.

Challenge 1272, page 788: Since the height of the potential is always finite, walls can always be overcome by tunnelling.

Challenge 1273, page 788: The lid of a box can never be at rest, as is required for a tight closure, but is always in motion, due to the quantum of action.

Challenge 1278, page 792: The one somebody else has thrown away. Energy costs about 10 cents/kWh. For new lamps, the fluorescence lamp is the best for the environment, even though it is the least friendly to the eye, due to its flickering.

Challenge 1279, page 796: This old dream depends on the precise conditions. How flexible does the display have to be? What lifetime should it have? The newspaper like display is many years away and maybe not even possible.

Challenge 1280, page 796: The challenge here is to find a cheap way to deflect laser beams in a controlled way. Cheap lasers are already available.

Challenge 1281, page 796: There is only speculation on the answer; the tendency of most researchers is to say no.

Challenge 1282, page 796: No, as it is impossible because of momentum conservation, because of the no-cloning theorem.

Challenge 1284, page 796: The author predicts that mass-produced goods using this technology (at least 1 million pieces sold) will not be available before 2025.

Challenge 1285, page 796: Maybe, but for extremely high prices.

Challenge 1286, page 798: For example, you could change gravity between two mirrors.

Challenge 1287, page 798: As usual in such statements, either group or phase velocity is cited, but not the corresponding energy velocity, which is always below c.

Challenge 1288, page 800: Echoes do not work once the speed of sound is reached and do not work well when it is approached. Both the speed of light and that of sound have a finite value. Moving with a mirror still gives a mirror image. This means that the speed of light cannot be reached. If it cannot be reached, it must be the same for all observers.

Challenge 1289, page 800: Mirrors do not usually work for matter; in addition, if they did, matter would require much higher acceleration values.

Challenge 1292, page 802: The overhang can have any value whatsoever. There is no limit. Taking the indeterminacy principle into account introduces a limit as the last brick or card must not allow the centre of gravity, through its indeterminacy, to be over the edge of the table.

Challenge 1293, page 802: A larger charge would lead to field that spontaneously generate electron positron pairs, the electron would fall into the nucleus and reduce its charge by one unit.

Challenge 1296, page 803: The Hall effect results from the deviation of electrons in a metal due to an applied magnetic field. Therefore it depends on their speed. One gets values around 1 mm. Inside atoms, one can use Bohr's atomic model as approximation.

Challenge 1297, page 803: The usual way to pack oranges on a table is the densest way to pack spheres.

Challenge 1298, page 804: Just use a paper drawing. Draw a polygon and draw it again at latter times, taking into account how the sides grow over time. You will see by yourself how the faster growing sides disappear over time.

Challenge 1299, page 805: The steps are due to the particle nature of electricity and all other moving entities.

Challenge 1300, page 805: Mud is a suspension of sand; sand is not transparent, even if made of clear quartz, because of the scattering of light at the irregular surface of its grains. A suspension cannot be transparent if the index of refraction of the liquid and the suspended particles is different. It is never transparent if the particles, as in most sand types, are themselves not transparent.

Challenge 1301, page 805: No. Bound states of massless particles are always unstable.

Challenge 1302, page 805: The first answer is probably no, as composed systems cannot be smaller than their own compton wavelength; only elementary systems can. However, the universe is not a system, as it has no environment. As such, its length is not a precisely defined concept, as an environment is needed to measure and to define it. (In addition, gravity must be taken into account in those domains.) Thus the answer is: in those domains, the question makes no sense.

Challenge 1303, page 806: Methods to move on perfect ice from mechanics:

- if the ice is perfectly flat, rest is possible only in one point – otherwise you oscillate around that point, as shown in challenge 21;

- do nothing, just wait that the higher centrifugal acceleration at body height pulls you away;
- to rotate yourself, just rotate your arm above your head;
- throw a shoe or any other object away;
- breathe in vertically, breathing out (or talking) horizontally (or vice versa);
- wait to be moved by the centrifugal acceleration due to the rotation of the Earth (and its oblateness);
- jump vertically repeatedly: the Coriolis acceleration will lead to horizontal motion;
- wait to be moved by the Sun or the Moon, like the tides are;
- 'swim' in the air using hands and feet;
- wait to be hit by a bird, a flying wasp, inclined rain, wind, lava, earthquake, plate tectonics, or any other macroscopic object (all objects pushing count only as one solution);
- wait to be moved by the change in gravity due to convection in Earth's mantle;
- wait to be moved by the gravitation of some comet passing by;
- counts only for kids: spit, sneeze, cough, fart, pee; or move your ears and use them as wings.

Note that gluing your tongue is not possible on perfect ice.

Challenge 1304, page 806: Methods to move on perfect ice using thermodynamics and electrodynamics:
- use the radio/tv stations to push you around;
- use your portable phone and a mirror;
- switch on a pocket lam, letting the light push you;
- wait to be pushed around by Brownian motion in air;
- heat up one side of your body: black body radiation will push you;
- heat up one side of your body, e.g. by muscle work: the changing airflow or the evaporation will push you;
- wait for one part of the body to be cooler than the other and for the corresponding black body radiation effects;
- wait for the magnetic field of the Earth to pull on some ferromagnetic or paramagnetic metal piece in your clothing or in your body;
- wait to be pushed by the light pressure, i.e. by the photons, from the Sun or from the stars, maybe using a pocket mirror to increase the efficiency;
- rub some polymer object to charge it electrically and then move it in circles, thus creating a magnetic field that interacts with the one of the Earth.

Note that perfect frictionless surfaces do not melt.

Challenge 1305, page 806: Methods to move on perfect ice using quantum effects:
- wait for your wavefunction to spread out and collapse at the end of the ice surface;
- wait for the pieces of metal in the clothing to attract to the metal in the surrounding through the Casimir effect;
- wait to be pushed around by radioactive decays in your body.

Challenge 1306, page 806: Methods to move on perfect ice using general relativity:
- move an arm to emit gravitational radiation;
- deviate the cosmic background radiation with a pocket mirror;
- wait to be pushed by gravitational radiation from star collapses;
- wait to the universe to contract.

Challenge 1307, page 806: Methods to move on perfect ice using materials science, geophysics, astrophysics:
- be pushed by the radio waves emitted by thunderstorms and absorbed in painful human joints;
- wait to be pushed around by cosmic rays;
- wait to be pushed around by the solar wind;
- wait to be pushed around by solar neutrinos;

- wait to be pushed by the transformation of the Sun into a red giant;
- wait to be hit by a meteorite.

Challenge 1308, page 806: A method to move on perfect ice using selforganisation, chaos theory, and biophysics:
- wait that the currents in the brain interact with the magnetic field of the Earth by controlling your thoughts.

Challenge 1309, page 806: Methods to move on perfect ice using quantum gravity, supersymmetry, and string theory:
- accelerate your pocket mirror with your hand;
- deviate the Unruh radiation of the Earth with a pocket mirror;
- wait for proton decay to push you through the recoil.

Challenge 1313, page 811: This is easy only if the black hole size is inserted into the entropy bound by Bekenstein. A simple deduction of the black hole entropy that includes the factor 1/4 is not yet at hand.

Challenge 1314, page 812: An entropy limit implies an information limit; only a given information can be present in a given region of nature. This results in a memory limit.

Challenge 1315, page 812: In natural units, the expression for entropy is $S = A/4 = 0.25A$. If each Planck area carried one bit (degree of freedom), the entropy would be $S = \ln W = \ln(2^A) = A \ln 2 = 0.693A$. This quite near the exact value.

Challenge 1319, page 816: The universe has about 10^{22} stars; the Sun has a luminosity of about 10^{26} W; the total luminosity of the visible matter in the universe is thus about 10^{48} W. A gamma ray burster emits up to $3 \cdot 10^{47}$ W.

Challenge 1324, page 818: They are carried away by the gravitational radiation.

Challenge 1336, page 839: Two stacked foils show the same effect as one foil of the same total thickness. Thus the surface plays no role.

Challenge 1338, page 841: The electron is held back by the positive charge of the nucleus, if the number of protons in the nucleus is sufficient, as is the case for those nuclei we are made of.

Challenge 1340, page 848: The number is small compare with the number of cells. However, it is possible that the decays are related to human ageing.

Challenge 1343, page 858: The nuclei of nitrogen and carbon have a high electric charge which strongly repels the protons.

Challenge 1344, page 861: Touching something requires getting near it; getting near means a small time and position indeterminacy; this implies a small wavelength of the probe that is used for touching; this implies a large energy.

Challenge 1345, page 868: Building a nuclear weapon is not difficult. University students can do it, and even have done so once, in the 1980s. The problem is getting or making the nuclear material. That requires either an extensive criminal activity or an vast technical effort, with numerous large factories, extensive development, coordination of many technological activities. Most importantly, such a project requires a large financial investment, which poor countries cannot afford. The problems are thus not technical, but financial.

Challenge 1348, page 881: Most macroscopic matter properties fall in this class, such as the change of water density with temperature.

Challenge 1351, page 892: Before the speculation can be fully tested, the relation between particles and black holes has to be clarified first.

Challenge 1352, page 893: Never expect a correct solution for personal choices. Do what you yourself think and feel is correct.

Challenge 1355, page 897: A mass of 100 kg and a speed of 8 m/s require 43 m^2 of wing surface.

Challenge 1358, page 907: The infinite sum is not defined for numbers; however, it is defined for a knotted string.

Challenge 1360, page 909: This is a simple but hard question. Find out.

Challenge 1334, page 822: No system is known in nature which emits or absorbs only one graviton at a time. This is another point speaking against the existence of gravitons.

Challenge 1363, page 912: Lattices are not isotropic, lattices are not Lorentz invariant.

Challenge 1365, page 913: Large raindrops are pancakes with a massive border bulge. When the size increases, e.g. when a large drop falls through vapour, the drop splits, as the central membrane is then torn apart.

Challenge 1366, page 913: It is a drawing; if it is interpreted as an image of a three-dimensional object, it either does not exist, or is not closed, or is an optical illusion of a torus.

Challenge 1367, page 913: See T. Fink & Y. Mao, *The 85 Ways to Tie a Tie*, Broadway Books, 2000.

Challenge 1368, page 913: See T. Clarke, Laces high, *Nature Science Update* 5th of December, 2002, or http://www.nature.com/nsu/021202/021202-4.html.

Challenge 1370, page 923: The other scale is the horizon of the universe, as we will see shortly.

Challenge 1371, page 924: Sloppily speaking, such a clock is not able to move its hands in such a way to guarantee precise time reading.

Challenge 1374, page 938: The final energy E produced by a proton accelerator increases with its radius R roughly as $E \sim R^{1.2}$; as an example, CERN's SPS achieves about 450 GeV for a radius of 740 m. Thus we would get a radius of more than 100 000 light years (larger than our galaxy) for a Planck energy accelerator. An accelerator achieving Planck energy is impossible.

A unification energy accelerator would be about 1000 times smaller. Nature has no accelerator of this power, but gets near it. The maximum measured value of cosmic rays, 10^{22} eV, is about one thousandth of the unification energy. The mechanisms of acceleration are obscure. Black holes are no sources for unification energy particles, due to their gravitational potential. But also the cosmic horizon is not the source, for some yet unclear reasons. This issue is still a topic of research.

Challenge 1375, page 938: The Planck energy is $E_{\text{Pl}} = \sqrt{\hbar c^5/G} = 2.0\,\text{GJ}$. Car fuel delivers about 43 MJ/kg. Thus the Planck energy corresponds to the energy of 47 kg of car fuel, about a tankful.

Challenge 1376, page 938: Not really, as the mass error is equal to the mass only in the Planck case.

Challenge 1377, page 938: It is improbable that such deviations can be found, as they are masked by the appearance of quantum gravity effects. However, if you do think that you have a prediction for a deviation, publish it.

Challenge 1379, page 939: There is no gravitation at those energies and there are no particles. There is thus no paradox.

Challenge 1381, page 939: The Planck acceleration is given by $a_{\text{Pl}} = \sqrt{c^7/\hbar G} = 5.6 \cdot 10^{51}\,\text{m/s}^2$.

Challenge 1382, page 940: All mentioned options could be valid at the same time. The issue is not closed and clear thinking about it is not easy.

Challenge 1383, page 940: The energy is the unification energy, about 800 times smaller than the Planck energy.

Challenge 1384, page 941: This is told in detail in the section on maximum force starting on page 985.

Challenge 1385, page 946: Good! Publish it.

Challenge 1386, page 952: See the table on page 171.

Challenge 1387, page 953: The cosmic background radiation is a clock in the widest sense of the term.

Challenge 1409, page 967: For the description of nature this is a contradiction. Nevertheless, the term 'universe', 'set of all sets' and other mathematical terms, as well as many religious concepts are of this type.

Challenge 1412, page 968: For concept of 'universe'.

Challenge 1416, page 969: Augustine and many theologians have defined 'god' in exactly this way. (See also Thomas Aquinas, *Summa contra gentiles*, 1, 30.) They claim that it possible to say what 'god' is not, but that it is not possible to say what it is. (This statement is also part of the official roman catholic catechism: see part one, section one, chapter one, IV, 43.)

Many legal scholars would also propose a different concept that fits the definition – namely 'administration'. It is difficult to say what it is, but easy to say what it is not.

More seriously, the properties common to the universe and to 'god' suggest the conclusion that the both are the same. Indeed, the analogy between the two concepts can be expanded to a proof. (This is left to the reader.) In fact, this might be the most interesting of all proofs of the existence of gods. This proof certainly lacks all the problems that the more common 'proofs' have. Despite its interest, the present proof is not found in any book on the topic. The reason is obvious: the result of the proof, the equivalence of 'god' and the universe, is a heresy for most religions.

If one is ready to explore the analogy nevertheless, one finds that a statement like 'god created the universe' translates as 'the universe implies the universe′. The original statement is thus not a lie any more, but is promoted to a tautology. Similar changes appear for many other – but not all – statements using the term 'god'. Enjoy the exploration.

Challenge 1417, page 969: If you find one, publish it! And send it to the author as well.

Challenge 1419, page 971: If you find one, publish it and send it to the present author as well.

Challenge 1423, page 975: Any change in rotation speed of the Earth would change the sea level.

Challenge 1424, page 976: Just measure the maximum water surface the oil drop can cover, by looking at the surface under a small angle.

Challenge 1425, page 977: Keep the fingers less than 1 cm from your eye.

Challenge 1426, page 983: As vacuum and matter cannot be distinguished, both share the same properties. In particular, both scatter strongly at high energies.

Challenge 1427, page 986: Take $\Delta f\ \Delta t \geqslant 1$ and substitute $\Delta l = c/\Delta f$ and $\Delta a = c/\Delta t$.

Challenge 1449, page 1026: The number of spatial dimensions must be given first, in order to talk about spheres.

Challenge 1450, page 1029: This is a challenge to you to find out and publish; it is fun, may bring success and would yield an independent check of the results of the section.

Challenge 1455, page 1040: The lid of a box must obey the indeterminacy relation. It cannot be at perfect rest with respect to the rest of the box.

Challenge 1456, page 1040: Of course not, as there are no infinite quantities in nature. The question is whether the detector would be as large as the universe or smaller. What is the answer?

Challenge 1461, page 1040: Yes, as nature's inherent measurement errors cannot clearly distinguish among them.

Challenge 1462, page 1040: Of course.

Challenge 1460, page 1040: No. Time is continuous only if *either* quantum theory and point particles *or* general relativity and point masses are assumed. The argument shows that only the combination of both theories with continuity is impossible.

Challenge 1463, page 1040: We still have the chance to find the best approximate concepts possible. There is no reason to give up.

Challenge 1464, page 1040: A few thoughts. beginning of the big bang does not exist, but is given by that piece of continuous entity which is encountered when going backwards in time as much as possible. This has several implications.

- Going backwards in time as far as possible – towards the 'beginning' of time – is the same as zooming to smallest distances: we find a single strand of the amoeba.
- In other words, we speculate that the whole world is one single piece, knotted, branched and fluctuating.
- Going far away into space – to the border of the universe – is like taking a snapshot with a short shutter time: strands everywhere.
- Whenever we sloppily say that extended entities are 'infinite' in size, we only mean that they reach the horizon of the universe.

In summary, no starting point of the big bang exists, because time does not exist there. For the same reason, no initial conditions for particles or space-time exist. In addition, this shows there was no creation involved, since without time and without possibility of choice, the term 'creation' makes no sense.

Challenge 1465, page 1040: The equivalence follows from the fact that all these processes require Planck energy, Planck measurement precision, Planck curvature, and Planck shutter time.

Challenge 1444, page 1008: The system limits cannot be chosen in other ways; after the limits have been corrected, the limits given here should still apply.

Challenge 1472, page 1064: Planck limits can be exceeded for extensive observables for which many particle systems can exceed single particle limits, such as mass, momentum, energy or electrical resistance.

Challenge 1476, page 1067: Do not forget the relativistic time dilation.

Challenge 1478, page 1068: Since the temperature of the triple point of water is fixed, the temperature of the boiling point is fixed as well. Historically, the value of the triple point has not been well chosen.

Challenge 1479, page 1068: Probably the quantity with the biggest variation is mass, where a prefix for $1\,\mathrm{eV}/c^2$ would be useful, as would be one for the total mass in the universe, which is about 10^{90} times larger.

Challenge 1480, page 1069: The formula with $n-1$ is a better fit. Why?

Challenge 1482, page 1071: No, only properties of parts of the universe. The universe itself has no properties, as shown on page 970.

Challenge 1483, page 1073: The slowdown goes *quadratically* with time, because every new slowdown adds to the old one!

Challenge 1484, page 1075: The double of that number, the number made of the sequence of all even numbers, etc.

Challenge 1477, page 1068: About 10 µg.

Challenge 1486, page 1077: This could be solved with a trick similar to those used in the irrationality of each of the two terms of the sum, but nobody has found one.

Challenge 1487, page 1077: There are still many discoveries to be made in modern mathematics, especially in topology, number theory and algebraic geometry. Mathematics has a good future.

Challenge 1488, page 1081: The gauge coupling constants determine the size of atoms, the strength of chemical bonds and thus the size of all things.

Challenge 1489, page 1095: Covalent bonds tend to produce full shells; this is a smaller change on the right side of the periodic table.

Challenge 1492, page 1100: $|z|$ is the *determinant* of the matrix $z = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$.

Challenge 1495, page 1100: Use Cantor's diagonal argument, as in challenge 1082.

Challenge 1498, page 1102: Any rotation by an angle 2π is described by -1. Only a rotation by 4π is described by $+1$; quaternions indeed describe spinors.

Challenge 1500, page 1105: Just check the result component by component. See also the mentioned reference.

Challenge 1502, page 1107: For a Gaussian integer $n + im$ to be prime, the integer $n^2 + m^2$ must be prime, and in addition, a condition on n mod 3 must be satisfied; which one and why?

Challenge 1505, page 1108: The metric is regular, positive definite and obeys the triangle inequality.

Challenge 1510, page 1110: The solution is the set of all two by two matrices, as each two by two matrix specifies a linear transformation, if one defines a transformed point as the product of the point and this matrix. (Only multiplication with a fixed matrix can give a linear transformation.) Can you recognize from a matrix whether it is a rotation, a reflection, a dilation, a shear, or a stretch along two axes? What are the remaining possibilities?

Challenge 1513, page 1111: The (simplest) product of two functions is taken by point-by-point multiplication.

Challenge 1514, page 1111: The norm $\|f\|$ of a real function f is defined as the supremum of its absolute value:

$$\|f\| = \sup_{x \in \mathrm{R}} |f(x)| . \qquad (851)$$

In simple terms: the maximum value taken by the absolute of the function is its norm. It is also called 'sup'-norm. Since it contains a supremum, this norm is only defined on the subspace of *bounded* continuous functions on a space X, or, if X is compact, on the space of all continuous functions (because a continuous function on a compact space must be bounded).

Challenge 1517, page 1116: Take out your head, then pull one side of your pullover over the corresponding arm, continue pulling it over the over arm; then pull the other side, under the first, to the other arm as well. Put your head back in. Your pullover (or your trousers) will be inside out.

Challenge 1521, page 1120: The transformation from one manifold to another with different topology can be done with a tiny change, at a so-called *singular point*. Since nature shows a minimum action, such a tiny change cannot be avoided.

Challenge 1522, page 1121: $M^\dagger M$ is Hermitean, and has positive eigenvalues. Thus H is uniquely defined and Hermitean. U is unitary because $U^\dagger U$ is the unit matrix.

So far, out of 1522 challenges, 669 solutions are given; in addition, 215 solutions are too easy to be included. Another 638 solutions need to be written; let the author know which one you want most.

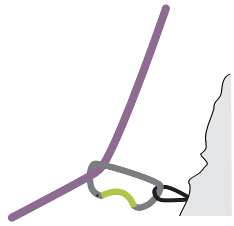
Appendix G

List of Illustrations

Picture credits

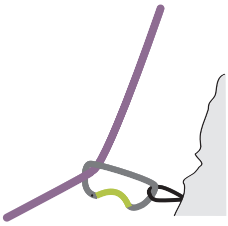
All drawings not explicitly marked are copyright © 1997 – 2005 by Christoph Schiller. The photograph on the cover is courtesy and copyright by Dave Thompson (http://www.daveontrek.co.uk). The figures on pages 31, 53 and 125 were made especially for this text and are copyright by Luca Gastaldi. The picture of milky water on page 533 was made for this text and is copyright by Antonio Martos. The copyrights of the figures on page 28 and on page 34 cannot be traced. The famous photograph of the Les Poulains and its lighthouse by Philippe Plisson on page 35 is courtesy and copyright by Pechêurs d'Images; see the websites http://www.plisson.com and http://www.pecheurs-d-images.com. It is also found in Plisson's magnus opus *La Mer*, a stunning book of photographs of the sea. The anticrepuscular photograph on page 60 is courtesy and copyright by Peggy Peterson. The firing caterpillar figure of page 63 is courtesy and copyright of Stanley Caveney. The measured graph of the walking human on page 88 is courtesy and copyright of Ray McCoy. The basilisk running over water on page 107 is courtesy and copyright by the Belgian group TERRA vzw (see http://www.terra.vzw.org). The water strider photograph on page 108 is courtesy and copyright by Charles Lewallen. The fluid flowing over an inclined plate on page 238 is courtesy and copyright by Vakhtang Putkaradze. The figure on the triangulation of the meridian of Paris on page 113 is copyright and courtesy of Ken Alder and found on his website http://www.kenalder.com. The analemma on page 128 and the Moon photographs on page 59 are courtesy and copyright by Anthony Ayiomamitis; the story of the photographs is told on his website at http://www.perseus.gr. The geoid of page 115 is courtesy and copyright by the GFZ Potsdam, to be found at http://www.gfz-potsdam.de. Figure 60 on page 106 is courtesy and copyright of the Australian Gemini project at http://www.ausgo.unsw.edu.au. The hollow Earth figure on pages 56 and 459 is courtesy of Helmut Diel and was drawn by Isolde Diel. The drawing of swirled spheres on page 234 is courtesy and copyright by Karsten Kötter. The photograph of a light pulse on page 252 is courtesy and copyright of Tom Mattick. The relativistic views on page 277 are courtesy and copyright of Daniel Weiskopf. The figure of myosotis on page 173 is courtesy and copyright by Markku Savela. The oscillon picture on page 234 is courtesy and copyright by Paul Umbanhowar. The figure of the soliton in the water canal on page 196 is copyright and courtesy of Dugald Duncan and taken from his website on http://www.ma.hw.ac.uk/solitons/soliton1.html. The figures of galaxies on page 403, pages 403 to 405 and on page 433 are courtesy of NASA. The Hertzsprung-Russell diagram on page 414 and the maps of the universe on page 411 are courtesy and copyright of Richard Powell, and taken from his website http://www.anzwers.org/free/universe. The pictures showing colour blindness on page 550 are courtesy and copyright of Michael Douma, from his splendid website at http://webexhibits.org/causesofcolor/2.html. The pictures of retinas on page 542 are courtesy and copyright of Austin Roorda. The ocean figure on page 501 is courtesy of Stefan Maus, and taken from his http://www.gfz-potsdam.de/pb2/pb23/SatMag/ocean_tides.html website. The picture of the rainbow on page 513 is from the NOAA website. The electroscope photograph on page 843 is courtesy and copyright of Harald Chmela and taken from his website http://www.hcrs.at. The simple motor photograph on page 502 is courtesy and copyright of Stefan Kluge. The photograph of lightning on page 482 is copyright S. Horsburgh (see http://www.horsburgh.com) and used with his permission. The pictures of snow flakes on page 804 are courtesy and copyright of Furukawa Yoshinori. The photograph of a lifter on page 500 is courtesy and copyright of Jean-Louis Naudin; more information can be found on his website http://www.jlnlabs.org. The supernumerary rainbow on page 518 is courtesy and copyright of Wolfgnag Hinz and from his website http://www.meteoros.de. The photograph of *M. bavaricum* on page 491 was made by Marianne Hanzlik and is courtesy of Nicolai Petersen. The MRI images of the head and the spine on page 837 are courtesy and copyright of Joseph Hornak and taken from his website http://www.cis.rit.edu/htbooks/mri/. The MRI image on page 839 of humans making love is courtesy and copyright of Willibrord Wejimar Schultz. The nuclide chart on page 842 is used with permission of Alejandro Sonzogni of the National Nuclear Data Center, using data extracted from the NuDat database. The pictures of feathers insects on page 899 are material used with kind permission of HortNET, a product of The Horticulture and Food Research Institute of New Zealand. The copyright of the figure on page 921 belongs to the M.C. Escher Heirs, c/o Cordon Art, Baarn, The Netherlands, who kindly gave permission for its use. The picture of the scallop on page 904 is courtesy and copyright of John Sullivan from the *Optiverse* website, on http://new.math.uiuc.edu/optiverse/. The picture of the butterfly in a wind tunnel on page 897 is courtesy and copyright of Robert Srygley and Adrian Thomas. The picture of the scallop on page 900 is courtesy and copyright of Dave Colwell.

Appendix H



List of Tables

Name Index



Page numbers in italic typeface refer to pages where the person is presented in more detail.

B

BLOCK

C

H

K
KARL

M

N

P

Penzias

Q

R

S

S

SALINGAROS

S

Sierra

W

WERTHEIM

X

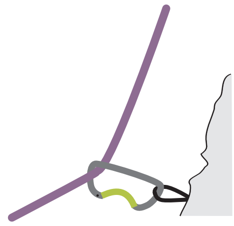
Y

Z

Z

ZBINDEN

Appendix J



Subject Index

Page numbers in ordinary typeface refer to pages where the keyword is used. Page numbers in *italic* typeface refer to pages where the keyword is defined or presented in detail. The subject index thus contains a glossary.

A

D

DIFFRACTION

E

EARTH-ALKALI

F

FAMILIARITY

G

H

I

I

IDENTITY

I

INVARIANT

L

LARGER

M

M

MANIFOLDS

M

MICROSCOPE

N

P

PARALLAX

P

PRAESEODYMIUM

Q

S

S

SATURN

S

SPECIAL

T

T

TECTONICS

U

U

UNIVERSAL

...

ma la religione di voi è qui

e passa

di generazione in generazione

ammonendo

che Scienza è Libertà.

Giosuè Carducci*

* '... but the religion of you all is here and passes from generation to generation, admonishing that science is freedom.' Giosuè Carducci (1835–1907), important Italian poet and scholar, received the Nobel Prize for literature in 1906. The citation is from Carducci's text inscribed in the entry hall of the University of Bologna, the oldest university of the world.