PHILOSOPHY OF PSYCHOLOGY

a contemporary introduction

José Luis Bermúdez

ROUTLEDGE CONTEMPORARY INTRODUCTIONS TO PHILOSOPHY

Philosophy of Psychology

"An outstanding introductory text in philosophy of psychology that lends itself readily to use in a variety of courses. It will, in addition, constitute an independent, substantive contribution to philosophy of psychology and philosophy of mind."

David Rosenthal, City University of New York, USA

"Philosophers of psychology and philosophically minded psychologists are in need of just this kind of introductory book. I would recommend this material both for pedagogy and as a place for scholars to turn to for a refresher."

Joe Cruz, Williams College, USA

Philosophy of Psychology is an introduction to philosophical problems that arise in the scientific study of cognition and behavior.

José Luis Bermúdez introduces the philosophy of psychology as an interdisciplinary exploration of the nature and mechanisms of cognition. He charts out four influential "pictures of the mind" and uses them to explore central topics in the philosophical foundations of psychology, covering all the core concepts and themes found in undergraduate courses in philosophy of psychology, including:

- Models of psychological explanation
- The nature of commonsense psychology
- Arguments for the autonomy of psychology
- Functionalist approaches to cognition
- Computational models of the mind
- Neural network modeling
- Rationality and mental causation
- Perception, action and cognition
- The language of thought and the architecture of cognition

Philosophy of Psychology: A Contemporary Introduction is a very clear and well-structured textbook from one of the leaders in the field.

José Luis Bermúdez is Professor of Philosophy and Director of the Philosophy-Neuroscience-Psychology Program at Washington University in St Louis, USA. He is a series editor of the International Library of Philosophy (Routledge) and author of *The Paradox of Self-Consciousness* (1998) and *Thinking without Words* (2003).

Routledge Contemporary Introductions to Philosophy

Series editor:
Paul K. Moser
Loyola University of Chicago

This innovative, well-structured series is for students who have already done an introductory course in philosophy. Each book introduces a core general subject in contemporary philosophy and offers students an accessible but substantial transition from introductory to higher-level college work in that subject. The series is accessible to non-specialists and each book clearly motivates and expounds the problems and positions introduced. An orientating chapter briefly introduces its topic and reminds readers of any crucial material they need to have retained from a typical introductory course. Considerable attention is given to explaining the central philosophical problems of a subject and the main competing solutions and arguments for those solutions. The primary aim is to educate students in the main problems, positions and arguments of contemporary philosophy rather than to convince students of a single position.

Classical Philosophy
Christopher Shields

Epistemology
Second Edition
Robert Audi

Ethics
Harry Gensler

Metaphysics
Second Edition
Michael J. Loux

Philosophy of Art
Noël Carroll

Philosophy of Language
William G. Lycan

Philosophy of Mind
Second Edition
John Heil

Philosophy of Religion
Keith E. Yandell

Philosophy of Science
Second Edition
Alex Rosenberg

Social and Political Philosophy
John Christman

Philosophy of Psychology
José Luis Bermúdez

Continental Philosophy
Andrew Cutrofello

Classical Modern Philosophy
Jeffrey Tlumak

Philosophy of Psychology

A contemporary introduction

José Luis Bermúdez

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2005
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2005.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

Library of Congress Cataloging in Publication Data
Bermúdez, José Luis.
Philosophy of psychology : a contemporary introduction / José Luis Bermúdez.
p. cm. – (Routledge contemporary introductions to philosophy)
Includes bibliographical references and index.
1. Psychology–Philosophy. I. Title. II. Series.
BF38.B46 2005

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-203-64240-6 Master e-book ISBN

ISBN 0-203-67593-2 (Adobe eReader Format)
ISBN 0-415-27594-6 (hbk)
ISBN 0-415-27595-4 (pbk)

Contents

Illustrations

Figures

Tables

Preface

I have written this book for advanced undergraduates and graduate students who wish to deepen their study of the mind by exploring central themes in the philosophy of psychology. The philosophy of psychology, as I see it, is the branch of philosophy focused primarily on the nature and mechanisms of cognition. How does thinking take place? What sort of representations does it involve? How should we understand transitions between those representations? How, if at all, are those transitions subject to criteria of rationality? Is a particular type of cognitive architecture required for cognition? Can we make any inferences from the nature and structure of high-level conscious thought to the nature and structure of the psychological mechanisms that underpin it?

The principal theme of this book is the interplay between the different ways of studying cognition and behavior in philosophy, scientific psychology and the neurosciences. The book explores how different conceptions of the mind operative in contemporary philosophy of psychology are grounded in different approaches to the scientific study of the mind. Chapter 2 presents the problem that I will be using to present these different approaches. This is what I call the *interface problem*. The interface problem is the problem of explaining how (if at all) commonsense (or folk psychological) explanations of mental states and behavior interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience, and the other levels in the hierarchy of disciplines devoted to the study of the mind/brain.

After Chapter 2 the book falls naturally into two parts. The next three chapters outline the principal "pictures" of the mind that emerge in response to the interface problem. Chapter 3 considers the pictures of the *autonomous mind* and the *functional mind*. According to the autonomy conception, there is a radical discontinuity between explanations given at the *personal level* of commonsense psychology and explanations given at the various *subpersonal* levels of explanation. The picture of the functional mind, in contrast, sees ordinary commonsense psychological explanations as a species of causal explanation, no more and no less mysterious than the various types of causal explanation with which we are familiar both from science and from our everyday experience of the physical world. The causal dimension of commonsense psychological explanation is what allows us to solve the interface problem. Proponents of the *representational mind* are motivated by a problem that they think cannot be tackled on a purely functional approach. This is the problem of *causation by content* – the problem of explaining how the

causal dimension of beliefs, desires and other mental states is a function of how they represent the world. The representational picture argues that this problem can only be solved by assuming that cognition takes place in a language-like representational medium. In this respect it is diametrically opposed to the fourth and final approach, which is the picture of the *neuro-computational mind*. This picture is strongly committed to the metaphor of the mind as brain and argues that our thinking about the mind must co-evolve with our thinking about the brain in a way that may lead to significant revisions of our commonsense ways of understanding cognition and behavior.

The final five chapters explore the dialectic between the four pictures of the mind in the context of specific issues and problems. These problems include how we should consider the causal dimension of psychological explanation (Chapter 6); how much of our social understanding and social interaction is underwritten by thinking about other people's behavior in terms of their beliefs and desires (Chapter 7); how we should think about the large-scale organization of the mind/brain (Chapter 8); whether appealing to notions of belief and desire in understanding cognition and behavior commits us to any specific hypotheses about physical structures in the brain (Chapter 9); and how we should understand the relation between thought and language (Chapter 10). The final chapter, Concluding thoughts, offers a speculative way of drawing together some of the strands that have emerged in the course of the book.

I am assuming that readers will already have a basic philosophical training and will have encountered some of the principal positions and arguments in the philosophy of mind. Technical philosophical terms and arguments are explained when they are first introduced, but readers will need some philosophical background to follow the explanations and ensuing discussion. Since my concern is primarily with the nature and mechanisms of cognition, there is relatively little discussion of the metaphysics of mind – of the philosophical questions that arise when one starts to think about the precise relations that might hold between mental states and brain states. There is considerable discussion of mental causation, but this is primarily in the context of the causal dimension of psychological explanation. It is not motivated by abstract metaphysical questions about how the domain of the mental can be accommodated with the realm of the physical. These are important questions, but not questions that are central to the philosophy of psychology. Nor am I primarily concerned with what are often termed theories of content. It is clear that mental states represent the world and that there are important questions to be asked about what fixes the particular way that a given mental state represents the world (just as there are important questions to ask about what fixes the particular way that a given spoken or written sentence represents the world). Philosophers have explored a number of different approaches to answering these questions. Some approaches stress causal relations between mental states and what they repre-

sent. Others are based on teleological theories of the function of mental states. I have prescinded from these debates, however, as they seem to me tangential to the questions that I am pursuing in this book. I take comfort in the thought that it is possible to make considerable progress in the philosophy of language without answering comparable questions about how words and sentences get their meaning.

At points in the text where philosophical questions that are not directly pursued become relevant I have tried to give guidance on recommended further reading. These recommendations will be found both in footnotes and in the annotated bibliography at the end of the book. The notes and bibliography will also help readers orient themselves in the enormous and often bewildering literature reporting and discussing the scientific study of the mind/brain. I have tried not to assume any background in psychology, cognitive science and neuroscience. Readers new to these topics should be able to follow and appreciate the significance of the examples, experiments and theories discussed. I hope that they will make use of the recommendations for further reading to deepen their knowledge of this fascinating interdisciplinary area.

Acknowledgments

The best audience on which to try out material for a textbook is a classroom of students and I have been fortunate to have had the opportunity to teach the material in this book to three classes of enthusiastic and critical students. I would very much like to thank the Spring 2002 *Philosophy of Psychology* class at Simon Fraser University in Vancouver; the Martinmas 2002 *Current Issues in the Philosophy of Mind* class in the St Andrews-Stirling graduate program; and the Fall 2003 *Philosophy of Psychology* class at Washington University in St Louis.

I owe a considerable debt to Gualtiero Piccinini for his detailed written comments on the penultimate draft, and to John Bickle for his helpful comments on Chapters 2 and 5. Santiago Amaya Gómez provided invaluable editorial and bibliographic assistance in putting together the final manuscript and preparing the index.

I am grateful to Tony Bruce for his initial invitation to contribute a volume to the Routledge Contemporary Introductions series and to Siobhan Pattison, Priyanka Pathak and Zoe Drayson for their patience as deadlines came and went. My thanks to Susan Dunsmore for her careful copy-editing.

Figure 1.1 taken from *Philosophical Writings of Descartes* (1985), Cambridge University Press, vol. 1, reproduced by permission of Cambridge University Press. Figures 2.1, 2.2 and Table 2.1 taken from D. Marr, *Vision* (1982), Henry Holt, reproduced by permission of Henry Holt & Company. Figure 3.2 adapted from T. Shallice, *From Neuropsychology to Mental Structure* (1988), Cambridge University Press, reproduced by permission of Cambridge University Press. Figures 5.1, 5.2 and 5.4 taken from P. McLeod, K. Plunkett and E. T. Rolls, *Introduction to Connectionist Modeling of Cognitive Processes* (1998), Oxford University Press, reproduced by permission of Oxford University Press. Figure 5.3 taken from J. L. Elman *et al.*, *Rethinking Innateness: A Connectionist Perspective on Development* (1996), MIT Press, reproduced by permission of MIT Press. Figures 6.1 and 6.2 taken from J. H. Holland, *Emergence: From Chaos to Order* (1998), Addison-Wesley, reproduced by permission of Pearson Education. Figure 8.1 taken from A. Ellis and A. Young, *Human Cognitive Neuropsychology* (1988), Psychology Press, reproduced by permission of Taylor & Francis Group. Figures 8.2 and 8.3 taken from R. A. Wilson and F. C. Keil (eds), *The MIT Encyclopedia of the Cognitive Sciences* (1999), MIT Press, reproduced by permission of MIT Press. Figure 10.1 taken from C. McDonald and G. McDonald, *Connectionism: Debates on Psychological Explanation* (1995), Blackwell, reproduced by permission of Blackwell Publishing Ltd.

1 What is the philosophy of psychology?

- What counts as psychology?
- Historical background
- Psychological concepts and the philosophy of psychology
- Philosophy of psychology and philosophy of mind

Many branches of philosophy are characterized as the philosophy of something else – from the philosophy of physics and the philosophy of economics to the philosophy of criticism. Broadly speaking, these all investigate the philosophical foundations of the relevant disciplines, exploring the high-level conceptual and empirical issues that cannot be tackled using the techniques and resources of those disciplines alone. The philosophy of psychology can also be described in these terms – as an investigation of the philosophical foundations of psychology. But the philosophy of psychology is distinctive, because the domain of investigation of the discipline whose foundations are being investigated overlaps with the domain of enquiry that philosophers have traditionally taken as their own. Philosophers have always taken it to be part of their brief to investigate the nature of mind and the nature of cognition. This sets up a parallelism of concern and corresponding scope for a two-way interaction that we do not find, for example, in the philosophy of economics or the philosophy of criticism. On the view developed in this book, the philosophy of psychology is the systematic study of the interplay between philosophical concerns and psychological concerns in the study of cognition. This interplay comes about because there are certain key concepts that feature both in the philosophical study of cognition and in the psychological study of cognition and that we cannot understand using the resources of either discipline on its own.

This introduction sketches out in more detail this guiding conception of the philosophy of psychology and provides some preliminary theoretical justification for it – the justification will be preliminary because the principal job will be carried out in the main body of the book. I start off in the first section with some comments about how I understand the domain of psychology. As will become apparent, I understand it very broadly indeed. In the second section I offer some examples of how blurred the boundaries were in the seventeenth and eighteenth centuries between what we would now think of as philosophical issues and psychological issues. In the third section I explain, and try to motivate, a particular view of the nature of theoretical concepts that underwrites the interactive conception of the philosophy of psychology. The fourth section explains what is distinctive about the philosophy of psychology as opposed to the philosophy of mind.

1.1 What counts as psychology?

A philosopher who undertakes to study the philosophy of economics will have a pretty clear sense of where to look to find their object of study – roughly speaking, the body of knowledge taught by, and the research carried out by, university economics departments and associated institutes and think tanks in the public and private sectors. Things are not so simple in the case of psychology. It is natural to think that psychology is the study of mind, behavior and the nature of cognition and action. But university psychology departments usually cover only a sub-set of the subject matters and disciplines that might intuitively be counted as psychological in this broad sense.

Many aspects of the investigation of the mind belong in medical faculties and/or hospitals. Cognitive neuropsychologists, for example, develop models of normal mental functioning by extrapolating from patterns of damage and impairment found in patients with neurological disorders.[1] What they do has not merely a theoretical dimension but also a directly diagnostic and therapeutic dimension. Similarly, much of our knowledge of the large-scale functioning of the brain in normal subjects comes from brain-imaging studies carried out on machines such as fMRI scanners whose primary function is diagnostic. As far as detailed knowledge of the fine-grained structure of the brain is concerned, almost everything that we know comes from neurophysiological experiments on animals employing techniques that, for example, allow scientists to record the activity of single neurons and to lesion identifiable neural areas. This is as much part of the general study of physiology as it is part of psychology. The same holds for much of our detailed knowledge of the nature of movement and action. Nor is all our knowledge of the mind and behavior experimental in origin. Observation pure and simple also has a role to play. An important contribution comes from the detailed observation of animals in the wild by cognitive ethologists and of infants and young children as they grow up by developmental psychologists. So too does cognitive modeling of the sort carried out by computer scientists, researchers into artificial intelligence and artificial life and computational neuroscientists.

In the face of this enormous range of disciplines and areas, I will for the purposes of this book make two stipulations: one exclusive and one inclusive. The exclusive stipulation is that I will not be considering much of what is done in psychology departments under the headings of social psychology or clinical psychology, in order to concentrate on the psychology of cognition and the related branches of the psychology of behavior. The inclusive stipulation is that I will treat as potentially relevant to the philosophy of psychology everything that bears upon the scientific study of cognition and behavior, whether it is carried out in psychology departments or not.

1 See Shallice (1988) for an influential overview of cognitive neuropsychology.

1.2 Historical background

The compartmentalization of psychological investigation is a relatively recent phenomenon. So too is the institutional separation of philosophy from the scientific study of cognition. The existence of psychology as an academic discipline goes back to the second half of the last century, and earlier than that work we might naturally think of as psychological was carried out by philosophers. In fact, even a brief look at the most important philosophers of the seventeenth and eighteenth centuries shows how deeply psychological much of their work was.

We can begin with Descartes, often cited as the father of modern philosophy. His philosophical enquiries into the possibility of knowledge were closely tied to his theory of what he took to be the actual workings of the mind. His view that our minds all contain a common core of innate ideas with which we are born is a crucial part of his explanation of how we can have knowledge of the world. Noam Chomsky, the best-known contemporary advocate of an innateness hypothesis, recognized the significance of Descartes's view by entitling one of his books *Cartesian Linguistics*. Similarly, Descartes saw clearly that his dualist theory that mind and matter are two fundamentally different types of thing demanded an account of how there could be interaction between them, and he developed a complicated theory based on the pineal gland to explain the workings of interaction. This theory included a sophisticated account of how the inverted images projected on the back of each retina were transmitted along nerve fibres to the brain and there reinverted and fused into an image on the retina (Figure 1.1). Descartes's work in these areas ranges effortlessly over what we now think of as the distinct areas of philosophy, psychology and physiology.

This lack of distinct disciplinary boundaries is equally clear in the three great British Empiricist philosophers: Locke, Berkeley and Hume. Both Locke's *An Essay Concerning Human Understanding* and Hume's *A Treatise of Human Nature* set out to provide a map of the range and scope of human knowledge. The distinctive feature of empiricist philosophy is the thought that all knowledge begins with the senses, in contrast for example to the Cartesian reliance on innate ideas and the independent functioning of reason. But developing this basic thought into a theory of knowledge requires explaining how the testimony of the senses can give rise to apparently non-sensory concepts and ideas. Consequently, both Locke and Hume provided psychological theories of how what they called complex ideas are generated from simple ideas through processes of abstraction, association, combination and comparison. Hume's theory in particular is the ancestor of much subsequent associationist theorizing, from behaviorist theories of conditioning to more recent research into neural network modeling. In both philosophers the limits they place on human knowledge and understanding are dictated by their psychological accounts of how ideas can be formed.

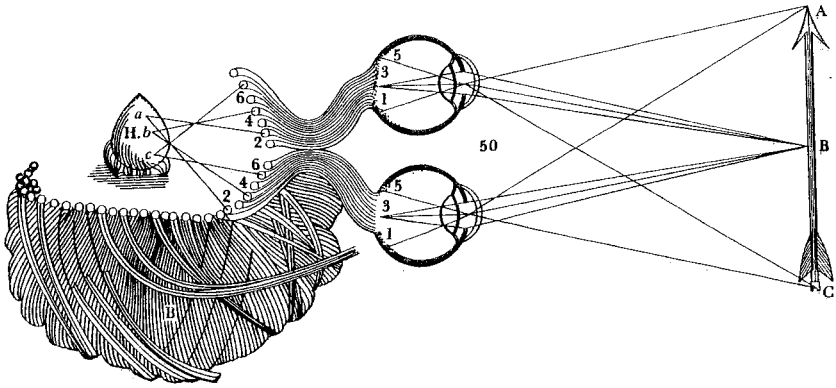

Figure 1.1 Descartes on the physiology of perception. Gland H. is the pineal gland (source: *Philosophical Writings of Descartes*, vol. 1 (1985)).

We find a different type of interplay between philosophical and psychological concerns in the writings of Bishop Berkeley. Berkeley famously defended an extreme form of idealism, according to which reality is a completely mental phenomenon. What we think of as physical objects are, according to Berkeley, collections of non-spatial ideas that exist in the mind of God when we are not perceiving them. Berkeley realized, of course, that this rather bizarre view cried out for an explanation of why it should seem so obvious to us that the world is made up *not* of collections of ideas, but rather of physical objects located in space. Since it is the evidence of our senses that provides the strongest support for the commonsense view of physical objects, Berkeley realized that his own theory had to be supported by a psychological account of sense perception, explaining how the 'illusions' of the commonsense view arise. This he provides, at least for the modality of vision, in two works, the *New Theory of Vision* and *The Theory of Vision Vindicated and Explained*.[2]

One obvious problem that Berkeley has to deal with is explaining why and how, if what we see doesn't really exist in space, we perceive what appears to be distance – where does the third dimension come from? Berkeley's answer to the problem of distance perception contains several ideas that were to become important in the psychology of vision. According to Berkeley, we do not perceive distance directly. What we perceive directly are two-dimensional ideas that contain what we would now call *cues* for distance – such as the sensation of turning one's eyes so that they are both aimed at the object; the blurred look that objects have when they are very close to the eyes; and the sensation of strain in the eyes that we have when we try to stop

2 Both of these are reprinted in M. R. Ayers's edition of Berkeley's selected writings (Berkeley 1975).

objects going out of focus as they approach our eyes. Berkeley's view is that these cues suggest an idea of distance ultimately derived from a conditioned association with the sense of touch (which for him includes what we would nowadays call *kinaesthesis* and *joint position sense*). He stressed, moreover, that the operation of these cues does not demand an unconscious inference – they work by something closer to association. Few commentators think that Berkeley did come up with a satisfactory theory of vision, but his attempt to formulate a psychology of vision that was compatible with his philosophy introduced ideas which were later to play an important part in the development of psychological thinking about vision, such as the concept of cues for distance and the idea that the spatial perception of distance and depth rests upon calibrating touch and vision (the idea that "touch educates vision").

As a final example of how blurred the boundaries were between psychology and philosophy in the days before the two disciplines were institutionally separated, it is worth having a brief look at Immanuel Kant. Although Kant is well known for his claim that there could never be a science of psychology because mental phenomena are not suitably quantitative, this claim should be interpreted with caution. There are, in Kant's view, several different types of psychology and at least one of them is well represented in his principal philosophical work, the *Critique of Pure Reason*. The type of psychology that Kant undertakes in the *Critique* is what he calls transcendental psychology. The purpose of transcendental psychology is to investigate (in a non-empirical, speculative manner) the general structural features that human cognition must have, given that it supports the type of experiences that it does.

So, for example, Kant is persuaded that the crucial cognitive activity must be what he calls *synthesis*, the bringing together of distinct representations under a single concept, and he argues that there are certain types of synthesis that underlie all the others. These basic types of synthesis, based on what he calls the pure concepts of the understanding, are the key to his analysis of human knowledge. They structure how we experience and interpret the world, and Kant argues that investigating them will explain the necessity and certainty of high-level principles, such as the principle that every event must have a cause. Admittedly, Kant's transcendental psychology is not based on empirical research; nor is it concerned to yield a bottom-level account of how cognition actually works. But this doesn't mean that it is not psychology. Rather, we should view much of what Kant says in the *Critique* as an exercise in psychology at what we would now call the computational level (Kitcher 1990; Brook 1994). Kant works from a specification of the cognitive tasks that human cognition must perform to a specification of the general features that any cognitive mechanism capable of performing those tasks must have.

As a more concrete illustration of this general point, consider Kant's discussion of spatial perception and Helmholtz's reaction to it (Hatfield 1990). Kant maintained that certain important features of our knowledge of space

could only be explained if space is in some sense innate. Kant attached particular significance to the idea that we are certain that space is Euclidean, and he argued that we could only have this certain knowledge if spatiality was something that we contributed to the world, rather than something that existed in the world independently of us. This line of argument seems clearly psychological. It prefigures later arguments for innateness hypotheses, such as those offered by Chomsky and Fodor, all of who argue for innateness as an explanation of how we can know things that we could not possibly have learnt. And it certainly imposes psychological obligations on those who dispute it. To deny, as Helmholtz was to deny, that space is innate requires showing that the kind of learning that Kant says is impossible really is possible. And this, of course, is what he tried to do, using experimental work on distorting prisms and newly-sighted patients to support a radically empiricist account of spatial perception. Kant's "philosophical" theory of the innateness of space throws out a psychological challenge that can be tested empirically. There seems no prospect of carving off the philosophical issues from the psychological issues.

This is not the place to explore how and why psychology and philosophy went their separate ways, fascinating story though this would be. More pressing is the question of how close the links between them *should* be. Was the move towards firm disciplinary boundaries and a clear division of labor a move in the right direction, or might something important have got lost in the professionalization of psychology and philosophy? We will explore this question in the next section.

1.3 Psychological concepts and the philosophy of psychology

An influential collection entitled *Essays in Conceptual Analysis* (Flew 1956) was published in the 1950s. It was intended to be a standard-bearer for a particular way of doing philosophy – the method of conceptual analysis. The guiding idea is that the business of philosophy is to analyze a range of central and fundamental concepts. The proper task of the philosophy of mind, for example, is to analyze such concepts as *belief*, *desire* and *intention*, while the central aim of epistemology is to provide an analysis of the concept of *knowledge*. Conceptual analyses are purely *a priori*. They are neither justified by nor answerable to any empirical facts that we might discover about the phenomena in question. They are obtained by reflecting on the connections between the various components of our conceptual scheme, by trying to identify relations of dependence between particular concepts and by constructing thought experiments that will test our intuitions and hence (so the theory goes) provide guidance as to how we understand particular concepts. So, for example, it was until quite recently the dominant conception of epistemology (the theory of knowledge) that it should proceed by constructing sets of necessary and sufficient conditions that

would pick out all and only the situations in which we would intuitively say that someone possessed knowledge. These necessary and sufficient conditions are "tested" by constructing hypothetical epistemic situations in which someone has a particular belief derived in a particular way, but where one or other condition is not satisfied, and then appealing to intuition to determine whether the belief in question really counts as knowledge.

If philosophy is purely a matter of conceptual analysis understood in this way, there is little scope for overlap between philosophy and psychology. For philosophers of the conceptual analysis school, our conceptual scheme has only the most tenuous of connections with empirical research in the natural or social sciences. Our everyday concept of perception, for example, is not in any way dependent upon research in the psychology of perception, nor will an analysis of our concept of knowledge involve any reference to the physiological and psychological mechanisms by which knowledge is actually acquired. Participation in the common conceptual scheme does not require scientific qualifications. So why should we need science to analyze the concepts within that scheme?

Few philosophers now think that this is the only way of doing philosophy – and even during the heyday of the conceptual analysis school there were many philosophers, particularly in North America, who had little sympathy with it.[3] Yet, even though obviously not a complete account of what philosophy is about, it does capture an important truth. Part of the job of philosophers is to explore and analyze the key concepts that we employ in thinking about ourselves and about the world. The problem with the conceptual analysis approach to philosophy is not with the basic idea that philosophers ought to analyze central concepts. It lies rather with how the conceptual analysis school understood the nature and aim of analysis.

As far as the aim of analysis is concerned, there is a certain futility in trying to find sets of necessary and sufficient conditions that will capture all and only the cases in which we would be disposed to apply concepts such as the concept of knowledge.[4] Debates about the validity or otherwise of proposed sets of necessary and sufficient conditions tend to center on complicated hypothetical cases to which our ordinary concepts may well not extend. Our ordinary conceptual scheme developed to provide a framework for thinking about the types of objects and situations that we tend to encounter,

3 Although a full history of twentieth-century analytic philosophy has yet to be written, I would conjecture that two key factors explain why the conceptual analysis model fared much better in the UK than in the USA. The first is that Quine's attack on the analytic/synthetic distinction was taken much less seriously in the UK. The second is the ascendancy in the UK of Wittgenstein and his followers. We will return to the first factor later in this section.

4 It is worth noting that many philosophers have moved away from the conceptual analysis approach to epistemology to what is known as naturalized epistemology, which sees the study of knowledge as continuous with the scientific investigation of the mechanisms by which we acquire knowledge. See the readings in Kornblith (1985), particularly the influential paper by Quine (Quine 1969). Kornblith (2002) defends the methodology of naturalized epistemology.

and we can expect it to be silent on such questions as whether or not to attribute knowledge to someone who finds himself in a region that he knows to be full of fake barn façades made from papier mâché and correctly identifies the object in front of him as a barn, even though he has not first checked to rule out the possibility that it might be a papier mâché barn façade.[5] The intuitions that philosophers canvas in discussing such hypothetical cases tend to reflect their prior theoretical commitments, rather than hidden depths of the concept purportedly under discussion. Reflection on this sort of case strongly suggests that any conception of conceptual analysis will only be workable if the constraints on what is to count as a successful analysis are relaxed. A conceptual analysis must be no more and no less imprecise and incomplete than the concept being analyzed – and if it is more precise and more complete (if it can be applied to situations for which our ordinary concepts are silent) then it should be recognized for what it is, namely, a refinement or sharpening of one of our everyday concepts, rather than an analysis of it.

But it is perfectly consistent to hold both that a successful conceptual analysis does not require necessary and sufficient conditions and that the business of conceptual analysis can proceed in complete independence of any empirical or scientific investigation. This is an influential view in contemporary philosophy (Lewis 1994; Jackson 1998). Yet it is in tension with two important insights into the nature of language and concepts that have been very influential in other areas of philosophy. One has become very well known, the other less so. Together they provide powerful reasons for thinking that the sorts of conceptual analysis undertaken in the philosophical study of the mind must be both informed by and responsive to empirical investigation of the mind.

The less well-known insight emerged during the prolonged discussion of Quine's attack on the analytic/synthetic distinction (Quine 1951). The idea that there is a sharp distinction between analytic truths, which are true in virtue of the meaning of the words they involve, and synthetic truths, which are true in virtue of the way the world is, is deeply implicated in the traditional conception of conceptual analysis – given the natural equation of concepts with the meanings of words. The truths revealed by successful conceptual analysis will be analytic truths. What stronger reason could there be for thinking that conceptual analysis can afford to ignore the empirical, given that empirical investigation can lead us only to synthetic truths?[6]

Hilary Putnam, although he did not agree with Quine that there was no distinction at all to be drawn between analytic truths and synthetic, took the

5 This is a famous epistemological example first put forward by Alvin Goldman in his paper, "Discrimination and Perceptual Knowledge" (Goldman 1976).

6 This close connection between the traditional conception of conceptual analysis and the analytic/synthetic distinction is one reason why traditional conceptual analysis has been more popular in the United Kingdom, where Quine's attacks on the analytic/synthetic distinction were never as widely accepted as they were in North America (due not least to the influential response in Grice and Strawson 1956).

view that the distinction was largely uninteresting (Putnam 1962). The only clear examples of analytic truths are relatively trivial, such as "all bachelors are unmarried men" or "a vixen is a female fox". Nor, on the other hand, should everything that does not count as analytic automatically be counted as synthetic. A synthetic statement (for Putnam) is one that can be confuted by isolated experiments or established by a process of enumerative induction. Many important statements fall into neither of these two categories. They have neither the stipulative and criterial character of genuine analytic statements nor the straightforwardly empirical character of genuine synthetic statements.

In drawing this general conclusion about the significance of the analytic/synthetic distinction Putnam drew our attention to an important category of concepts – what he termed *law-cluster concepts*. Many theoretical and scientific concepts are identified by the laws in which they feature. The concept of kinetic energy is what it is simply in virtue of the laws explaining how kinetic energy is created, preserved and transformed into other types of energy. These laws fix the meaning of the expression 'kinetic energy' and by so doing fix the identity of the concept *kinetic energy*. Putnam stresses that relatively few concepts have their identities fixed by a single law. Most scientifically interesting concepts are what he calls law-cluster concepts:

> The concept 'energy' is a great example of a law-cluster concept. It enters into a great many laws. It plays a great many roles, and these laws and inference roles constitute its meaning collectively not individually. I want to suggest that most of the terms in highly developed science are law-cluster concepts, and that one should always be suspicious of the claim that a principle whose subject term is a law-cluster term is analytic. The reason it is difficult to have an analytical relationship among law-cluster concepts is that such a relationship would be one more law. But, in general, any one law can be abandoned without destroying the identity of the law-cluster concept involved, just as a man can be irrational from birth, or have a growth of feathers all over his body, without ceasing to be a man.
>
> (Putnam 1962, p. 52)

Principles and statements involving law-cluster concepts fall into the gray area between the clear-cut analytic and the clear-cut synthetic. On the one hand, they are not criterial of the meaning of the law-cluster in the way that it is criterial of the concept *bachelor* that it apply only to unmarried men. On the other, they are too general and abstract to be overturned by isolated experiments.

This notion of a law-cluster concept provides a model for thinking about some key psychological concepts – in particular those that straddle the boundary between philosophy and psychology. I am thinking here of concepts such as *rationality*, *perception*, *cognition*, *reasoning*, *information*, *representation*, *understanding*, *action* and, of course, the concept *concept* itself. These

concepts, and others like them, feature in both philosophical and psychological discussion of cognition. I would suggest that these concepts, integral to the philosophy of psychology, should be identified in terms of all the different roles they play in different levels of theorizing about the mind. They are cluster concepts that cannot properly be understood unless one explores the full range of theories in which they feature – from the tacit and implicit theory of commonsense psychology that many theorists think that we all deploy to navigate the social world to the empirical studies of cognitive psychologists and the mathematical models developed by computational neuroscientists. It would be no less of a mistake to think that the resources of commonsense psychology will tell us everything we need to know about, say, the concept of rationality than it would be to think that rationality can be completely understood through empirical studies of people's reasoning habits. A proper understanding of the concept will come only through integrating the different strands in the cluster.

There are two significant differences between the notion of a cluster concept that I am suggesting applies to these central psychological concepts and Putnam's notion of a law-cluster concept. First, the cluster concepts explored in the philosophy of psychology are not best viewed as *law*-cluster concepts. Even if one thinks that commonsense psychology is theory-like and hence is law-like in some form or other, there are very few laws in psychology (Patterson 1996; Cummins 2000). It would be more appropriate to describe concepts such as the concept of rationality of the concept of consciousness as *theory-cluster concepts*. Psychology features many types of explanatory theory that do not involve laws.

Second, Putnam's law-cluster concepts (such as *kinetic energy* or *gravitational mass*) are much more clearly delineated. The concept *kinetic energy* features in many different laws, but they are all closely related and part of physics. Nothing like this is true of theory-cluster concepts such as *rationality* or *representation*. These concepts feature both in our commonsense conceptual scheme and in the scientific study of cognition. Even within the scientific study of cognition they feature at various different levels of explanation. We find representations appealed to both in discussions of personal-level conscious decision-making and in discussions of subpersonal-level cognitive processing.[7] We find them discussed in the context of language-processing and also attributed to non-linguistic creatures. The connection between these different uses and theories is far from clear. The challenge of the philosophy of psychology is to work towards a unified and integrated account of concepts such as these.

Since the key concepts investigated in the philosophy of psychology are theory-cluster concepts the activity of the philosophy of psychology can be characterized both as conceptual analysis and as essentially interdisciplinary

7 The distinction between personal and subpersonal levels of explanation is explored in more detail in Chapter 2. See particularly section 2.2.

and scientifically informed. Nonetheless, there is a naturally occurring worry at this point. It concerns our ordinary, pre-scientific understanding of such key concepts as, say, *rationality* and *perception*. Surely, one might think, we all learn these concepts and employ them in our pre-theoretical understanding of ourselves and others. We would not be able to do this unless we had an adequate understanding of them. Yet, how can we have an adequate understanding of them if a proper analysis of those concepts requires us to investigate the complexities of scientific psychology, cognitive science and the neurosciences? Surely, one might think, the process of conceptual analysis must be a process of making explicit what is implicit in our everyday concept mastery and concept use, and it is hard to see how anything that requires detailed scientific investigation can be implicit in our everyday concepts.

This brings us to the second insight mentioned earlier. Theorists have moved away from the idea that anybody who uses a linguistic term properly and with understanding must have a full grasp of the meaning of the sort that could be developed into a satisfying theoretical account of the associated concept. On our simplifying assumption that concepts should be understood as the meanings of the corresponding words, what has been rejected is a version of the idea that conceptual analysis can only reveal what is implicit in the ordinary, competent use of those concepts. According to *semantic externalism*, which grew out of ideas initially put forward by Hilary Putnam (Putnam 1975) and Tyler Burge (Burge 1982) the psychological states of a competent language user are not sufficient to fix the meaning of an important class of linguistic expressions. The meaning of these terms is partially fixed by the nature of the external environment. The thesis of semantic externalism has been worked out in most detail for so-called *natural kind* terms (terms, such as 'water' or 'gold', that, as the saying goes, "carve nature at its joints" by picking out the independently specifiable categories into which objects in the world fall). Putnam's original claim was that the meaning of a natural kind term includes the objects of which it is true (its extension) and that the particular "stereotype" that a speaker attaches to a word may well serve to latch on to characteristic exemplars of the type in question but will typically not be sufficient to determine whether problematic cases fall within the term's extension. Ordinary language-users will need to defer to experts for arbitration on these difficult issues. These experts will typically operate with criteria for determining the extension of words/concepts that are not familiar to ordinary language-users/concept-possessors. Here, then, we have a precedent for the idea that we should look to experts rather than to ordinary concept users for theoretical elucidation of our central psychological concepts.

It is a further implication of semantic externalism in the philosophy of language that, before the development of science made available techniques for determining the extension of natural kind terms/natural kind concepts, language users/concept users could have been systematically mistaken about the extension (and hence about the meaning) of these externally

individuated terms. Here is Putnam making the point with reference to 'gold'. He is discussing a piece of metal that is superficially similar to gold (it falls under the stereotype of gold) but as a matter of fact is not gold, although this can be detected only with modern techniques. Suppose that an ancient Greek had come across this piece of metal and classified it as gold. Would he have been mistaken?

> In the view I am advocating, when Archimedes asserted that something was gold (χρυσος) he was not just saying that it had the superficial characteristics of gold (in exceptional cases, something may belong to a natural kind and *not* have the superficial characteristics of a member of that natural kind, in fact): he was saying that it had the same *hidden structure* (the same 'essence', so to speak) as any normal piece of local gold. Archimedes would have said that our hypothetical piece of metal X was gold, but he would have been *wrong*.
>
> (Putnam 1975, pp. 235–236)

One implication is that it is perfectly possible for a community of concept possessors to make systematic and undetectable errors about the nature of a concept, simply because they do not have a deep enough scientific understanding of the phenomena that concept picks out. Furthermore, it is perfectly possible for something's "hidden essence" to be at odds with the stereotype through which we, as ordinary concept possessors, identify it.

Both points are very relevant to the philosophy of psychology. Scientific psychology, cognitive science and cognitive neuroscience are all in their infancy, as their practitioners would be the first to admit. It is perfectly possible that we are in the same position with reference to the concepts that define the domain of the philosophy of psychology as Archimedes might have been with gold or a seventeenth-century natural philosopher with the concept of force. Perhaps what we take to be the definitive nature of the concept *rationality*, manifest to us in everyday thought and communication, stands to the real nature of rationality in something like the way the stereotype attaching to the concept *gold* stands to the "hidden essence" of gold – a set of beliefs and preconceptions that allow us to latch on to a genuine cognitive phenomenon but that we should not assume will be the last word in analyzing that phenomenon.[8] Of course, we cannot simply abandon the conception of rationality implicit in our everyday conceptual scheme – or at least not without very good reason. But we must not forget that the obligation of answerability goes in two directions. Our scientific

8 Something like this has been suggested for the case of belief by William Lycan: "As in Putnam's examples of 'water', 'tiger' and so on … the ordinary word 'belief' (*qua* theoretical term of folk psychology) points dimly towards a natural kind that we have not fully grasped and that only a mature psychology will reveal" (1988, p. 32). Some of the consequences of this view of psychological vocabulary are worked out in the special case of knowledge in Kornblith (2002).

investigations must be sensitive to our pre-theoretical understanding of the concepts in question, but so too must we be prepared to change our pre-theoretical understanding in response to what we learn from empirical investigation.

In the case of the natural kind terms discussed by Putnam, the division of labor (to use his own phrase) is relatively clear. We, as ordinary concept possessors and language users, have readily identifiable experts to whom we can defer when we are unsure about whether or not to apply a concept in a particular situation. There is little serious dispute about whom we should consult or where we should go to find out whether something is gold or not – or whether a tree is an elm or a beech. But in the case of the philosophy of psychology things are not so simple. There is no settled conception of whom we should defer to when we are trying to apply and understand our core psychological concepts. As we will discover in the next chapter, different ways of thinking about the mind identify different ultimate authorities. Two extreme views can easily be identified. Some philosophers will be unimpressed by everything I have so far said in this chapter and will insist that our tacitly understood commonsense psychological concepts must be the ultimate court of appeal. This yields what in the next chapter I will characterize as the *autonomous* conception of the mind. Other philosophers, and many neuroscientists, will think that we should defer to the findings of neuroscience. This is the conception that I will term the *neurocomputational* conception of the mind. No doubt the truth lies somewhere between these extremes, and we will explore this dialectic further in subsequent chapters.

For the moment, the point to extract is simply that the interactive conception of the philosophy of psychology can be grounded quite plausibly in an account of psychological concepts as *theory-cluster concepts*. The philosophy of psychology is in the business of conceptual analysis, but not in the business of conceptual analysis of the standard *a priori* variety. Theory-cluster concepts require investigation that is both conceptual in the standard sense and empirical. The challenge for the theorist trying to analyze a theory-cluster concept is to integrate the different strands of the cluster – to construct an integrated account out of what appears to be a single concept occurring in seemingly incommensurable theories. Breakthroughs are made when it turns out that apparently incommensurable theories are not really incommensurable after all – when a way is discovered of integrating theories at different levels of description, for example. But, conversely, the constant danger is that what appears to be a single concept is not really a single concept after all – when it turns out that theorists at different levels of description are using similar words to express radically different concepts.

1.4 Philosophy of psychology and philosophy of mind

In order to fix more clearly what the philosophy of psychology is, it will be useful to explain how I see it differing from the philosophy of mind. This is

not an area in which it is possible to draw a sharp dividing line since both branches of philosophy are obviously concerned with the mind in a broad sense. Yet they are concerned with the mind in different ways, and the differences are differences of substance rather than emphasis, even though, as one would expect, the two branches of philosophy are deeply complementary.

Many of the issues that dominate the philosophy of mind have to do with the metaphysics of the mind – with how we are to categorize the mind and its states in ontological terms. Textbooks and courses in the philosophy of mind typically begin by discussing the attractions and drawbacks of dualism and then go on to discuss the alternatives to dualism that have been canvassed in the philosophical literature – various forms of the identity theory, functionalism, eliminative materialism, and so on. Discussion then typically moves on to how, if at all, it is possible for the mind to have a causal impact on the world. Again the emphasis is primarily metaphysical. The point at issue is how the mind fits into the world. Other central problems and topics in the philosophy of mind have a more epistemological dimension, most obviously the problem of other minds (the problem of explaining the grounds of our beliefs about the mental states of other people) but also the problem of explaining the distinctive character of our access to the contents of our own minds.

In contrast to these metaphysical and epistemological preoccupations the concerns of the philosophy of psychology are more directly focused on the activity of cognition and on the explanation of behavior. How does cognition take place? What sort of representations does it involve? How should we understand transitions between those representations? How, if at all, are they subject to criteria of rationality? Is a particular type of cognitive architecture required for cognition? Can we make any inferences from the nature and structure of high-level conscious thought to the nature and mechanisms of the psychological mechanisms that underpin it? These are typical questions in the philosophy of psychology that will recur throughout this book and that are clearly distinct from the metaphysical and epistemological questions predominating in the philosophy of mind.

Whatever position one takes on the details of dividing up the intellectual terrain between the philosophy of mind and the philosophy of psychology (and different authors will do it in different ways), it seems clear that there is a broad methodological divergence between the two branches. This divergence concerns the scope for interdisciplinarity. To the extent that the guiding problems in the philosophy of mind are metaphysical and epistemological in nature, there will be little need in tackling them to go into much empirical detail. So, for example, no amount of neurophysiological and neuropsychological research establishing neural correlates for conscious personal-level psychological states could possibly entail the truth of the claim that psychological states are identical to brain states. The existence of correlations between mental states and brain states is compatible with every

position on the metaphysical nature of those states, and nothing that one might say about the metaphysics of the mind is empirically refutable. No self-respecting dualist would want to rule out the possibility, for example, that there might be neural correlates for non-physical mental states. Indeed, the most plausible contemporary version of dualism, the property dualism propounded by David Chalmers, incorporates a program for studying the physical correlates and counterparts of non-physical phenomenal properties (Chalmers 1996).

In summary, then, the philosophy of psychology (as I understand it and as I will be presenting it in this book) differs from the philosophy of mind in two basic ways (although we should view these differences as shifting positions relative to each other on a continuum, rather than as sharp qualitative distinctions). First, the philosophy of psychology is concerned primarily with the nature and mechanisms of cognition, rather than with the metaphysics and epistemology of the mind. Second, and as a direct consequence of the previous point, the philosophy of psychology lacks the insulation from scientific research and concerns that more traditional debates in the philosophy of mind possess in virtue of their metaphysical and epistemological dimension.

2 Levels of psychological explanation and the interface problem

- Explanation at different levels
- Personal and subpersonal levels of explanation
- Horizontal explanation, vertical explanation and commonsense psychology
- The interface problem and four pictures of the mind

We can study a living organism as a collection of particles, as a dynamical system, as a structure with a complex chemical composition, as a biological entity, as a part of an ecosystem, and so on. To each of these ways of looking at a living organism there corresponds a distinct and often self-standing level of explanation. Many, perhaps most, scientists believe that there is some order in this multiplicity of perspectives, that the different levels of explanation can be linked together to yield a unified account of the living organism.

In the special case where the living organism has a mind, there is a range of further ways of characterizing it. We might describe it in terms of the cognitive functions it can perform (perceiving, for example, or calculating a long division sum). Or we might talk about the cognitive mechanisms that allow it to perform those functions. Alternatively, we might talk about the physical structure within which those cognitive mechanisms are to be found. To each way of talking there corresponds a distinct psychological level of explanation and, as with the non-cognitive levels of explanation, the dream and the hope of many researchers are that a unified account bringing together all these levels of explanation will eventually be forthcoming. This chapter explores the intuitively appealing idea that these different levels of explanation come together in a hierarchical structure.

In section 2.1 the general idea of explanation at different levels is developed in more detail, taking as a case study David Marr's analysis of the visual system, one of the most significant theoretical achievements of recent psychology and one whose guiding idea is that psychological explanation takes place at different levels. As we will see, there are limitations to Marr's conception of how different levels of explanation mesh together. In section 2.2 I introduce the important distinction between personal-level and subpersonal-level states. Personal-level states are states of the thinking and acting organisms and they feature in a distinctive type of explanation of the behavior of such organisms. Section 2.3 develops this conception of personal-

level explanation in more detail, outlining the widely held view that explanation at the personal level involves explaining and predicting the behavior of cognitive agents in terms of *commonsense psychology*. As we see in section 2.4, this suggestion leads naturally to what I term the *interface problem*. This is the problem of explaining the relation between the commonsense, everyday type of psychological explanation that we all engage in every day (or so at least it is claimed) and the levels of explanation lower down in the hierarchy. How do explanations of the behavior of people given in terms of their beliefs, desires and other psychological states mesh, for example, with explanations in terms of patterns of activity across populations of neurons? How does the biochemistry of what goes on inside a neuron relate to the dynamics of how a person interacts with the environment? What is the relation between understanding a person as a conscious, reasoning agent, on the one hand, and understanding that person's brain as a complicated type of computational mechanism? In section 2.5 I provide a brief overview of four different ways of responding to the interface problem. These four responses yield the four different pictures of the mind that we will use as a thread to explore the philosophy of psychology.

2.1 Explanation at different levels

The mind can be studied at many different levels. We can study the mind from the bottom up, beginning with individual neurons and populations of neurons, or perhaps even lower down, with molecular pathways whose activities generate action potentials in individual neurons, and then trying to build up from that by a process of *reverse engineering* to higher cognitive functions (reverse engineering being the process by which one takes an object and tries to work backwards from its structure and design to the function it performs). Or we can begin from the top down, starting out with general theories about the nature of thought and the nature of cognition and working downwards to investigate how corresponding mechanisms might be instantiated in the brain. On either approach one will proceed via distinct levels of explanation that often have separate disciplines corresponding to them.

The idea that these different levels form a clearly defined hierarchy is well established among those who write about the theoretical dimension of psychology and cognitive science. Daniel Dennett, for example, distinguishes between explanation from the *intentional* stance at the top of the hierarchy, beneath which is explanation from the *design* stance and then explanation from the *physical* stance (Dennett 1987). At the intentional stance we consider a system (which could be a human agent, a cognitive system such as the memory system, or an artifact such as a chess computer) as if it were a rational thinking agent attempting to solve a particular task or set of tasks. We identify the constraints that such a task imposes and the general strategy or strategies that it might employ to solve those tasks. When we adopt

the design stance we move down a level to consider the general principles and constraints governing the design of a system that might solve those tasks. Going down a step further we move to the physical stance where we consider how a system with the appropriate sort of design might actually be physically constructed. In the study of human cognition, for example, it is when we adopt the physical stance that we have to come to terms with the constraints imposed by the physical structure of the brain.

A broadly similar tripartite distinction can be found in the model of the human visual system developed by David Marr (Marr 1982). Marr's model of the visual system is the best-worked-out analysis of how different levels of explanation can be combined in the elucidation of a cognitive phenomenon. The approach that Marr took to linking levels of explanation has been deeply influential, both among practicing scientists and among philosophers interested in understanding the nature of psychological explanation. Although, as we shall see in more detail in the next three chapters, there are several different and competing conceptions of how such links might work, it will be useful to start with Marr to get a general flavor of how a single theoretical account might straddle several different levels of explanation.

Marr distinguishes three different levels at which the visual system can be analyzed. The top level is the *computational level*, dealing with the general constraints posed by the particular type of task that is being carried out. The task of an analysis at the computational level is (a) to translate a general description of the cognitive phenomenon in which we are interested into a specific account of a particular information-processing problem that is being solved; and (b) to identify the constraints within which any solution to the information-processing task must operate. The guiding assumption here, of course, is that cognition is ultimately to be understood in terms of information-processing – in terms of processes that transform one kind of information (say, the information coming into a cognitive system through its sensory systems) into another type of information (say, information about what type of objects there might be in the organism's immediate environment). A computational analysis will identify the information with which the cognitive system has to begin (the *input* to that system) and the information with which it needs to end up (the *output* from that system).[1]

The next step down in understanding how the visual system works comes with what Marr calls the *algorithmic level*. Research at the algorithmic level takes the form of specifying a detailed set of information-processing instructions that will be able successfully to solve the information-processing problem identified at the computational level. The essence of any information-processing task is the transformation of a given input into a given output. The input could be information from the sensory systems about the

1 The basic principles of the information processing approach to cognition will become clearer in the following, but more detail and useful background information will be found in Crane (1995, Ch. 3) and in Harnish (2002).

Computational theory	*Representation and algorithm*	*Hardware implementation*
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 2.1 Three levels at which a system carrying out an information-processing task can be understood (source: Marr (1982)).

distribution of light in the visual field, or it could be a description of the layout of the pieces on a chessboard. Correspondingly the output might be a three-dimensional representation of the environment around a perceiver, or a proposed move within a game of chess. The main task at the algorithmic level is to specify a way of representing both input and output that will allow the formulation of an algorithm (a series of computational steps, similar to those undertaken by a calculator) to transform input into output. In contrast, the principal task at the *implementational level* is to find a physical realization for the algorithm – that is to say, to identify physical structures that will realize the representational states over which the algorithm is defined and to find mechanisms at the neural level that can properly be described as computing the algorithm in question (Figure 2.1).

The approach Marr proposes is a paradigm example of what is called *top-down* analysis. He starts with high-level analysis of the specific information-processing problems that the visual system confronts, as well as the constraints under which the visual system operates. At each stage of the analysis these problems become more circumscribed and more determinate. The suggestions offered at the algorithmic and implementational levels are motivated by discussions of constraint and function at the computational level – that is, by considering which features of the environment the organism needs to model and the resources it has available to it.

In thinking about the general functioning of the visual system and the constraints under which any account operates Marr leant heavily on research on brain-damaged patients carried out by clinical neuropsychologists. In his book *Vision* (Marr 1982), he explicitly refers to Elizabeth Warrington's work on patients with damage to the left and right parietal cortex – a type of brain damage typically associated with deficits in perceptual recognition. Warrington noticed that the perceptual deficits of the two classes of patient are fundamentally different. Patients with right parietal lesions are able to recognize and verbally identify familiar objects *provided that they can see them from familiar or "conventional" perspectives*. From unconventional perspectives,

however, these patients would not only fail to identify familiar objects but would also vehemently deny that the shapes they perceived could possibly correspond to the objects that they in fact were. Patients with left parietal lesions showed a diametrically opposed pattern of behavior. Although left parietal lesions are often accompanied by language problems, patients with such lesions tend to be capable of identifying objects (as manifested in successful performance on matching tasks).

From this pattern of breakdown Marr drew two conclusions about how the visual system functions (following a standard, but not uncontroversial, pattern of inference from the existence of *dissociations* between cognitive abilities in brain-damaged patients to the conclusion that those abilities are subserved by different forms of information processing in the brain).[2] He concluded, first, that information about the shape of an object must be processed separately from information about what those objects are for and what they are called and, second, that the visual system can deliver a specification of the shape of an object even when that object is not in any sense recognized. Here is Marr describing how he used these neuropsychological data to work out the basic functional task that the visual system performs:

> Elizabeth Warrington had put her finger on what was somehow the quintessential fact about human vision – that it tells us about shape and space and spatial arrangement. Here lay a way to formulate its purpose – building a description of the shapes and positions of things from images. Of course, that is by no means all that vision can do; it also tells us about the illumination and about the reflectances of the surfaces that make the shapes – their brightnesses and colors and visual textures – and about their motion. But these things seemed secondary; they could be hung off a theory in which the main job of vision was to derive a representation of shape.
>
> (Marr 1982, p. 7, cited in Cummins and Cummins 1999, p. 79)

So, at the functional level, the basic task of the visual system is to derive a representation of the three-dimensional shape and spatial arrangement of an object in a form that will allow that object to be recognized. Since ease of recognition is correlated with the ability to extrapolate from the particular vantage point from which an object is viewed, Marr concluded that this description of object shape should be on an object-centered rather than an egocentric frame of reference (where an egocentric frame of reference is one

2 The guiding assumption behind this type of inference from brain damage to mental structure has been termed the assumption of *subtractivity* (Saffran 1982), namely, that the performance of a neuropsychological patient reflects total normal cognitive functioning minus those systems that have been impaired (rather than the operations of new post-traumatic brain structures). For a clear presentation of the role that the subtractivity assumption plays in cognitive neuropsychology, see Shallice (1988). Martha Farah has raised some important theoretical issues about the methodology of cognitive neuropsychology (Farah 1994). See also Caramazza (1986).

centered on the viewer). This, in essence, is the theory that emerges at the computational level.[3]

Moving to the algorithmic level, clinical neuropsychology drops out of the picture and the emphasis shifts to the very different discipline of psychophysics – the experimental study of perceptual systems. When we move to the algorithmic level of analysis we require a far more detailed account of how the general information-processing task identified at the computational level might be carried out. Task-analysis at the computational level has identified the type of inputs and outputs with which we are concerned, together with the constraints under which the system is operating. What we are looking for now is an algorithm that can take the system from inputs of the appropriate type to outputs of the appropriate type. This raises a range of new questions. How exactly is the input and output information encoded? What are the system's *representational primitives* (the basic "units" over which computations are defined)? What sort of operations is the system performing on those representational primitives to carry out the information processing task?

A crucial part of the function of vision is to recover information about the reflectance, distance and orientation of visible surfaces. In Marr's theory this information is derived from a series of increasingly complex and sophisticated representations, which he terms the *primal sketch*, the *2.5D sketch* and the *3D sketch*. At the algorithmic level the job is to specify these different representations and how the visual system gets from one to the next, starting with the basic information arriving at the retina. Since the retina is composed of cells that are sensitive to light, this basic information is information about the intensity of the light reaching each of those cells. In thinking about how the visual system might work, we need (according to Marr) to think about what properties of the retinal information might provide clues for recovering the information we want about surfaces and their reflectance, distance, orientation, and so forth. What are the starting-points for the information-processing that will yield as its output an accurate representation of the lay-out of surfaces in the distal environment? Marr's answer is that the visual system needs to start with discontinuities in light intensity, because these are a good guide to boundaries between objects and other physically relevant properties. Accordingly the representational primitives that he identifies are all closely correlated with changes in light intensity. These include *zero-crossings* (registers of sudden changes in light intensity), blobs, edges, segments and boundaries. The algorithmic description of the visual system takes a representation formulated in terms of these representational primitives as the input, and endeavors to spell

3 Of course, the functional specification of the visual system is not purely top-down and derived from high-level disciplines such as cognitive neuropsychology. The job of the visual system is to compute a representation of three-dimensional shape on the basis of the fundamental inputs that it receives. Marr characterizes these fundamental inputs as the changes in intensity values at specific points in the visual array that are detected by photoreceptors in the retina and passed into the visual system via the lateral geniculate nucleus. Clearly, therefore, physiological information is playing a role in the task-analysis of the visual system.

Table 2.1 Representational framework for deriving shape information from images

Name	*Purpose*	*Primitives*
Image(s)	Represents intensity.	Intensity value at each point in the image.
Primal sketch	Makes explicit important information about the two-dimensional image, primarily the intensity changes there and their geometrical distribution and organization.	Zero-crossings Blobs Terminations and discontinuities Edge segments Virtual lines Groups Curvilinear organization Boundaries
$2\frac{1}{2}$-D sketch	Makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame.	Local surface orientation (the "needles" primitives) Distance from viewer Discontinuities in depth Discontinuities in surface orientation
3-D model representation	Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives (i.e., primitives that represent the volume of space that a shape occupies) as well as surface primitives.	3-D models arranged hierarchically, each one based on a spatial configuration of a few sticks or axes, to which volumetric or surface shape primitives are attached.

Source: Marr (1982)

out a series of computational steps that will transform this input into the desired output, which is a representation of the three-dimensional perceived environment (Table 2.1).

We can work through a single example to get a better sense of the sort of questions that arise in thinking about how to spell out these computational steps. A crucial stage in visual processing is working out the orientation of visible surfaces. There is an important question to be settled here about how the visual system represents and calculates surface orientation (see Marr 1982, §3.7). Traditional accounts of vision have assumed that surface orientation is computed from texture gradients. The texture gradient of a surface is the way in which the fineness of detail that can be seen in it decreases in direct proportion to increasing distance from the observer. A cobbled street is a classic example. The cobbles up close are sharply defined and clearly identifiable, but as they get further away the smoother they appear. Texture gradient is an important cue for depth and much exploited by visual artists. The evidence from psychophysics, however, is that surface orientation is

represented in terms of the coordinates of slant and tilt. Slant is the angle by which a perceived surface falls away from the frontal (i.e. the vertical) plane, while tilt is the direction of the slant. If you stand a book on the table in front of you and move the top backwards/forwards you are altering its slant, while if you move one side backwards/forwards you are changing its tilt. But in constructing an algorithm to compute surface orientation, one needs to determine how the visual system represents the extent of slant and the extent of tilt. It might do so in terms of angles, or perhaps in terms of ratios between the lengths of the sides of the triangle whose apex is the perceiver and whose base is the surface in question – e.g. the sine, cosine or tangent of

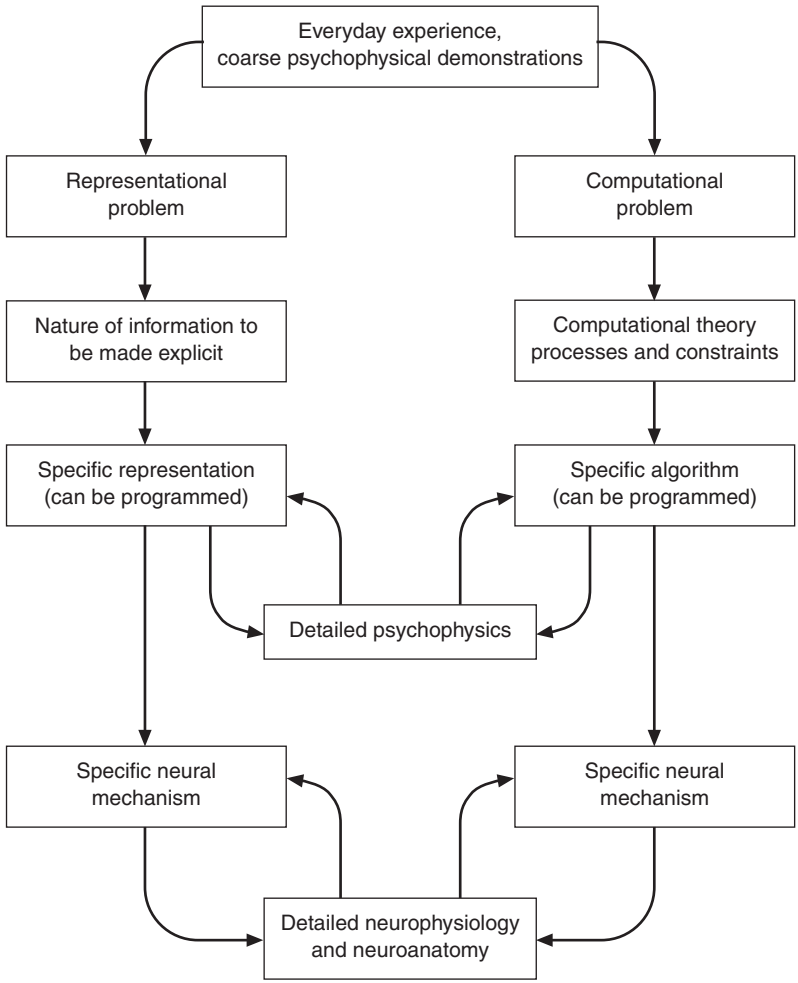

Figure 2.2 Relationships between representation and processes (source: Marr (1982, p. 332)).

those angles. Useful clues come from psychophysics. We can infer the quantities in terms of which the extent of slant and tilt are being computed by working backwards from the relation between the errors that subjects make when judging surface orientation in a range of different conditions. It turns out that there is a uniform rate of error correlated with the angles of slant and tilt rather than to any function related to those angles. The natural conclusion to draw is that the visual system is sensitive to angles directly rather than to ratios between lengths, and this will need to be reflected in the algorithm developed to compute surface orientation.

Moving down to the implementational level a further set of disciplines come into play. In thinking about the cognitive architecture within which the various algorithms computed by the visual system are embedded we will obviously need to take into account the basic physiology of the visual system – and this in turn is something that we will need to think about at various different levels. Marr's own work on vision contains relatively little discussion of neural implementation. But Figure 2.2 illustrates where the implementational level fits into the overall picture, according to Marr.

Marr's analysis of the visual system, therefore, gives us a clear illustration not only of how a single cognitive phenomena can be studied at different levels of explanation, but also of how the different levels of explanation can come together to provide a unified analysis. Marr's top-down approach clearly defines a hierarchy of explanation, both delineating the respective areas of competence of different disciplines and specifying ways in which those disciplines can speak to each other. It is not surprising that Marr's analysis of the visual system is frequently taken to be a paradigm of how scientific psychology ought to proceed. But Marr's particular version of the hierarchical conception is in one respect of very limited application. It does not pretend to be even a complete account of vision. It only deals with what is sometimes called *early visual processing* – that is to say, the visual processing that parses the visual array into three-dimensional objects standing in certain spatial relation to each other. But in many ways this is only the beginning of an account of vision. An analysis of early visual processing will have little to tell us about the more complex dimensions of visual perception – such as, for example, how perceptual recognition works; how the way we see the world allows us to act within it and upon it; how we perceive motion and distinguish our own motion from the motion of objects; how we coordinate our visually-derived picture of the world with information from the other senses and from the various somatic feedback systems telling us about bodily position and orientation. Still less will it tell us how we are able to remember things that we have seen or how we come to a decision about what to do on the basis of what we see.

It has become common among psychologists and cognitive scientists to draw a distinction between modular and non-modular cognitive processes.[4]

4 The classic presentation of the distinction between modular and non-modular processing is Fodor (1983).

This is, in essence, a distinction between high-level cognitive processes that are open-ended and involve bringing a wide range of information to bear on very general problems, and lower-level cognitive processes that work quickly to provide rapid solutions to highly determinate problems. In more detail, modular processes are generally held to have most, if not all, of the following characteristics:

- *Domain-specificity*. They are highly specified mechanisms with a relatively circumscribed functional specification and field of application.
- *Mandatory application*. They respond automatically to stimuli of the appropriate kind, rather than being under any executive control.
- *Fast*. They transform input (e.g. patterns of intensity values picked up by photoreceptors in the retina) into output (e.g. representations of three-dimensional objects) quickly enough to be used in the on-line control of action.
- *Informational encapsulation*. Modular processing remains unaffected by what is going on elsewhere in the mind. Modular systems cannot be "infiltrated" by background knowledge and expectations.
- *Fixed neural architecture*. It is often possible to identify determinate regions of the brain associated with particular types of modular processing.
- *Specific breakdown patterns*. Modular processing can fail in highly determinate ways (as we saw in Marr's discussion of Elizabeth Warrington's patients). These breakdowns can provide clues as to the form and structure of that processing.

We will return to the distinction between modular and non-modular processing in subsequent chapters (particularly in Chapter 8). For the moment, we can simply note two things. First, the early visual system appears to be almost a paradigm of a modular system. Second, there seem to be very close relations between applicability to the early visual system of a Marr-style top-down analysis and its modularity.

The key to Marr's particular version of the top-down approach to the study of cognitive processes is that a suitable analysis at the functional level will yield a determinate task or set of tasks that it is the job of the cognitive system to perform. It is certainly true that, *at some level of generality*, even non-modular cognitive processes can be described as performing a particular function. But the point of task-analysis at the functional level is that the function or functions identified must be circumscribed and determinate enough for it to be feasible to identify an algorithm to compute them, and it is not obvious how this might be achieved for non-modular systems. It is relatively easy to see how the right sort of functional analysis might emerge when we are dealing with a cognitive process that is domain-specific and specialized – the task of functional analysis is essentially the task of clarifying what exactly the system is specialized to do.

A second relevant point is that algorithms must be computationally tractable. It must be possible to implement them in an organism in a way that will yield useful results within the appropriate time frame (which might be very short when it comes, for example, to predator detection). If an algorithm is to be specified, then there must only be a limited number of representational primitives and possible parameters of variation. Once again, it is easy to see why informational encapsulation will secure computational tractability. An informationally encapsulated module will have only a limited range of inputs on which to work (although there are important questions about how this filtering process is supposed to work; see section 8.4 below). In contrast, non-modular processing runs very quickly into versions of the so-called *frame problem* (Dennett 1984; Pylyshyn 1984). This is the problem, particularly pressing for those developing expert systems in AI and designing robots, of building into a system rules that will correctly identify what information and which inferences should be pursued in a given situation. The problem is identifying what sort of information is relevant and hence needs to be taken into account. Dennett's classic article on the subject opens with the following amusing and instructive tale:

> Once upon a time there was a robot, named R1 by its creators. Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT (Wagon, Room, t) would result in the battery being removed from the room. Straightaway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 knew that the bomb was on the wagon in the room, but didn't realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act.
>
> Back to the drawing board. "The solution is obvious," said the designers. "Our next robot must be made to recognize not just the intended implications of its acts, but also the implications about their side-effects, by deducing these implications from the descriptions it uses in formulating its plans." They called their next model, the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT (Wagon, Room, t) it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the colour of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon – when the bomb exploded.

> Back to the drawing board. "We must teach it the difference between relevant implications and irrelevant implications," said the designers, "and teach it to ignore the irrelevant ones." So they developed a method of tagging implications as either relevant or irrelevant to the project at hand, and installed the method in their next model, the robot-relevant-deducer, or R2D1 for short. When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting, Hamlet-like, outside the room containing the ticking bomb, the native hue of its resolution sicklied o'er with the pale cast of thought, as Shakespeare (and more recently Fodor) has aptly put it. "Do something!" they yelled at it. "I am," it retorted. "I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and …" the bomb went off.

The greater the range of potentially relevant information, the more intractable this problem will be. Conversely, the problem is unlikely to arise for a system that is informationally encapsulated in Fodor's sense – an informationally encapsulated module has built into it a solution to the frame problem.

Of course, it is hard to see how one might go about *proving* that top-down analysis fitting Marr's general model is only possible when one is dealing with systems that are modular in the strict Fodorean sense. But it should be clear that nothing like Marr's account could be *straightforwardly* applied to what we might think of as higher (i.e. non-modular) cognitive processes. So, it can hardly serve as a template for understanding how different levels of explanation might form a hierarchy. Moreover, even if it could be extended to non-modular processes, it would still fall a long way short of providing a picture of the mind as a whole. Whether or not it is possible to provide a functional specification susceptible to algorithmic formulation for high-level cognitive processes, it will certainly be impossible to do so for the mind as a whole – and it is, of course, an understanding of the mind as a whole that we are ultimately aiming for. Marr's analysis of the early visual system provides a clear illustration of the general idea of a hierarchy of different levels of explanation. But it is not itself pitched at the right sort of level to provide a model of how we might understand the general idea of a hierarchy of explanation applied to the mind as a whole. In the next section we will start to look in more detail at how to formulate the problem.

2.2 Personal and subpersonal levels of explanation

The general idea of a hierarchical conception is, as we have seen, a natural way of dealing with the fact that the study of the mind is carried out by a great range of academic disciplines, each with their own specialized aims and specialized techniques. If just one of these disciplines is the "right" way of approaching the mind, then it looks as if the others will end up dropping

out of the picture. But the dominant conception of the relation between the different ways of approaching the scientific study of the mind is much more tolerant and ecumenical. Although the serious scientific study of the mind is still in its infancy, in comparison with the scientific study of the non-sentient parts of the physical world, the guiding conception is that the different disciplines will eventually slot together to give a unified pyramid-like conception of the mind, just as it is often believed that the natural sciences slot together to give a unified, multi-level explanatory picture of the physical world. Many philosophers, psychologists and cognitive scientists think that we need to see the different disciplines as operating at different levels of the hierarchy, offering explanations that complement rather than compete with each other.

The basic idea behind the hierarchical approach to the study of the mind is that each different level elucidates the level above it. The agenda for the whole hierarchy, therefore, is set by the level of explanation at the top of the hierarchy. We saw in the previous section that, when we are thinking about the mind as a whole, there are difficulties applying the type of functional analysis that Marr applied to the early visual system. When we are thinking about the mind as a whole it is very difficult, and perhaps even impossible, to identify tasks that can be understood in a determinate enough way to yield algorithms. Let us take a different approach, moving away from functional analysis to explore the type of explanation that stands at the top of the hierarchy.

A very natural suggestion is that the top level of explanation must deal with the explanation and prediction of behavior. Cognition is not an isolated activity and if we are interested in studying the mind as a whole we must start from the twin facts, first, that it is organisms that have minds and, second, that possessing a mind allows those organisms to behave in the ways characteristic of intelligent agents. The top level of explanation deals with the mind as a whole and it is natural to think that we cannot do this without considering how cognitive agents behave. Theories such as Marr's operate at a lower level than the level of cognitive agents. They deal with parts or modules of the cognitive agent, rather than with the agent itself as a thinking and acting organism. They are theories at the subpersonal level (below the level of the person). It is natural to think, however, that what we want at the top level of the hierarchy of explanation is a theory that deals with the thinking and acting person.

We shall look in more detail in the next section at the form a personal-level theory will take, but for the moment we will simply concentrate on the distinction between personal and subpersonal states. The point of the personal–subpersonal distinction is not to collapse together all the different levels of explanation below commonsense psychology into a single subpersonal level of explanation. There are, of course, many different levels of subpersonal explanation – including almost all of what we think of as cognitive science and scientific psychology, as well as cognitive neuroscience, neurobi-

ology, and so forth. The real point, rather, is that there is a systematic ambiguity in our psychological vocabulary that can prevent us from correctly identifying what lies at the top level of the hierarchy. We can explore this ambiguity through two examples.

Many philosophers and psychologists place considerable stress on the cognitive significance of possessing a *cognitive map*, where the notion of a cognitive map is defined as a way of representing the spatial relations between things that is independent of the thinker's own spatial location – as opposed, for example, to representing spatial relations relative to a frame of reference centered on one's own body (Eilan *et al.* 1993). A subject who possesses a cognitive map can think about space independently of his own trajectory through it. Possession of a cognitive map in this sense is often thought to be a vital element in a subject's understanding of the objectivity of the spatial environment, and indeed of his being self-conscious (Campbell 1994). This nexus of ideas ultimately goes back to Kant's *Critique of Pure Reason*. In this first sense of cognitive map, possession of a cognitive map is a high-level cognitive ability, something whose attainment in childhood marks a significant ontogenetic step. It is a form of personal-level knowledge: knowledge of the spatial layout of a mind-independent world.

But there is another important sense in which the notion of a cognitive map is deployed. In this second sense, cognitive maps refer to the storage of geometric information in the nervous system. Here is a recent definition from Gallistel:

> A cognitive map is a record in the central nervous system of macroscopic geometric relations among surfaces in the environment used to plan movements through the environment.
>
> (Gallistel 1990, p. 103)

As with the first sense of 'cognitive map', we are dealing here with the simultaneous representation of spatial relations. But the suspicion that these spatial relations are not being represented in the same way is confirmed when we read on in Gallistel's *The Organization of Learning* and find that all animals from insects upwards possess similar types of cognitive maps in this second sense – the cognitive maps that control movement in animals all preserve a system of metric relations within earth-centered coordinates. This is clearly something very different from the first sense of 'cognitive map'. And it is a difference that one might capture by saying that 'cognitive map' is a personal-level term when used in the first sense, and a subpersonal-level term when used in the second sense.

As a second example, consider the state of looking at a particular object – say, a horse – and recognizing what sort of an object it is. The concepts that I possess lead me to classify that perceived object in a certain way. The result is a perceptual belief that I see a horse. There is a superficial similarity with David Marr's theory of visual information processing. According to Marr,

the final stage of visual information processing involves associating shape descriptions derived from the visual image with stored shape descriptions and 3-D models. This is also a form of visual classification, but apparently not of the same type as the other. Classification of the second type can occur without classification of the first type (just as one can have a cognitive map of the second kind without a cognitive map of the first kind). This is precisely the sort of difference that might be characterized by saying that the first state (the conscious recognitional state) is a personal-level state, while the second state (the state of the visual processing system) is a subpersonal-level state.

Examples such as these can give an intuitive grasp on the personal/subpersonal distinction, but it would be helpful to have criteria for picking out personal-level states. Several such criteria have been put forward:

1 *Accessibility to consciousness*. This has been pressed by John Searle (Searle 1990b). This criterion has obvious appeal for those who think that consciousness is the mark of the mental – and it seems true that any conscious or potentially conscious state is a personal-level state. But the converse does not appear to hold. There seem to be several types of personal-level states that would fail to qualify if accessibility to consciousness were the criterion. One example is the strongly unconscious states that feature in the psychological explanations offered in psychodynamic therapy (where, unlike a dispositional belief, a strongly unconscious state can remain in principle inaccessible to consciousness). Such psychoanalytic explanations seem to have many commonalities with paradigm instances of personal-level explanations and it would be unfortunate to rule them out as a matter of definition.[5] Another example comes from the tacitly known states implicated in language mastery. These seem inaccessible to consciousness. Even if one came consciously to believe a principle that one in fact employs in the grammatical analysis of heard utterances (perhaps after closely studying transformational linguistics) this would still not be to access the principle itself. Yet for many philosophers, understanding one's language seems a paradigmatically personal-level phenomenon.
2 *Cognitive penetrability*. This is the criterion proposed by Pylyshyn (1980). A state is cognitively penetrable if it is rationally sensitive to the subject's propositional attitudes (i.e. their beliefs, desires, hopes, fears and so forth). What this means is that a cognitively penetrable state will alter in response to relevant changes in a subject's beliefs, desires and other propositional attitudes (on the assumption, of

5 Of course some explanation needs to be given of why a psychological state should be strongly unconscious and some explanations using the concept of repression seems to imply a degree of awareness of the state in question, arguably implying a form of accessibility to consciousness. But it is implausible that all explanations in this area will take this form. For further discussion of this issue, see Gardner (1993). The thesis that psychodynamic explanations are personal-level is explored in Gardner (2000).

course, that propositional attitudes are canonical personal-level states). There are two major problems with this. First, the notion of rational sensitivity is far from clear. It is a mistake to think that some sort of relation of inferential integration holds across the whole set of personal-level states. How could it? They are almost all in long-term memory and long-term memory is only ever partially searched. Second, there seem some very clear counter-examples to the idea that cognitive penetrability is a necessary condition for personal-level states. Perceptual illusions are obviously personal-level states, but it is very well known that they are not cognitively penetrable. Knowing that the two lines are the same length in the Müller–Lyer illusion doesn't stop one looking longer than the other.

3 *Inferential integration*. One might modify the requirement of cognitive penetrability by suggesting that a personal-level state is *either* rationally sensitive to paradigm propositional attitudes *or* such that paradigm propositional attitudes are rationally sensitive to it. This would be to say that personal-level states are inferentially integrated with the body of a subject's propositional attitudes. This avoids the second of the two problems with cognitive penetrability, because paradigm propositional attitude states are clearly rationally sensitive to perceptual states. Nonetheless, the first problem still stands, since the notion of rational sensitivity remains central. Moreover, the earlier difficulties posed by tacitly known principles of language comprehension and strongly unconscious states remain in play, because it is doubtful whether there is rational sensitivity in either direction between either of these two categories and the main part of a subject's propositional attitude system.

It looks, therefore, as if none of these proposed criteria can on its own demarcate the realm of the personal level – and nor, of course, should this be very surprising. Hardly any concepts of theoretical interest can be captured within the scope of a neat set of necessary and sufficient criteria. It is true, nonetheless, that the disjunction of the three proposed criteria is a useful tool for picking out personal-level states – we can be pretty confident that any personal-level state will be *either* accessible to consciousness, *or* cognitively penetrable *or* inferentially integrated. But, as one would expect from a disjunction, it tells us little about the real nature of personal-level states. For that we would, I think, be better advised to look at the explanatory role that such states are called upon to play. This is what will occupy us in the next section.

2.3 Horizontal explanation, vertical explanation and commonsense psychology

As we shall see, each of the four pictures of the mind that we will be considering starts off from a particular conception at the personal level of how the mental states and thinking behavior of cognitive agents are to be explained.

Although they do not all understand explanation at the top of the hierarchy in quite the same way, the really fundamental differences between them come when we ask about the connections that hold between explanation at the top of the hierarchy and explanation at lower levels. In the final section of this chapter I give an overview of these four different pictures of the mind. Before doing that, however, it will be useful to work out a theoretical framework that will allow the differences between these four different conceptions to emerge in full focus. I shall start by introducing an important distinction between two different types of psychological explanation (*horizontal explanation* and *vertical explanation*).

Horizontal explanation is the explanation of a particular event or state in terms of distinct (and usually temporally antecedent) events or states. Horizontal explanations are singular and dated. That is, they specify relations between individual and identifiable events holding at a particular time. The paradigm is singular causal explanation – the explanation of the causal antecedents of a particular event. Suppose we ask why the window broke when it did. A horizontal explanation of the window's breaking might cite the baseball's hitting it, together with a generalization about windows tending to break when hit by baseballs travelling at appropriate speeds. Similarly, if we ask why the dendrite fired when it did, a horizontal explanation might cite the more or less simultaneous arrival of two nerve impulses at the synapse of an adjacent neuron, together with a generalization about the power of their combined potentials to evoke a spike potential in the adjacent dendrite.[6]

However, we can ask why-questions to which horizontal explanations are not appropriate answers. And we can continue to ask why-questions even when a horizontal explanation has been given. I can ask why the window broke when the baseball hit it, or why the combined potentials of the two neurons should have evoked a spike potential in the dendrite. In neither case will I be satisfied by having repeated to me the generalization that windows tend to break when baseballs hit them or that a certain combined potential in neurons firing almost simultaneously will tend to evoke a spike potential in a suitably placed dendrite. What I want to know is *why* those generalizations hold. I want to find out what features of the physical structure of glass make it the case that windows are fragile enough to be broken by baseballs – or about how chemical neurotransmitters induce new post-synaptic potentials in neurons. Of course, there are certain basic laws for which it is inappropriate to ask why they hold. Explanation must run out somewhere – but

6 The notion of horizonal explanation is intended to be more general than the deductive-nomological model of explanation proposed in Hempel and Oppenheim (1948). There is no requirement, for example, that a successful explanation should show that the phenomenon to be explained is a logical consequence of antecedent events in the light of the relevant generalizations. And many successful horizontal explanations will deploy generalizations that would doubtless not count as law-like by Hempel and Oppenheim's lights.

not with either generalizations about how windows behave or generalizations about how neurons behave.

Explanations given in response to these second types of why-question are *vertical explanations*. The project of vertical explanation can broadly be characterized as explaining the grounds of horizontal explanations. Differing conceptions of the appropriateness of vertical explanations will be generated by different conceptions of the sorts of grounds required by different types of horizontal explanation. It is vertical explanatory relations that hold between different levels of explanation. Typically, when questions of vertical explanation are asked, they are answered at a lower level of explanation. So different conceptions of vertical explanation will go with different conceptions of the relations between the different levels of the explanatory hierarchy.

With this ground-clearing behind us we can move on to the first of the questions identified earlier. What sort of horizontal explanations lie at the top of the hierarchy? It is widely believed that at the top of the hierarchy of psychological explanation there lies a form of psychological explanation of intelligent behavior that has been given various names – *commonsense psychology*, *folk psychology*, *theory of mind, naïve psychology* etc. (In the following I shall talk primarily of commonsense psychology.) There are different conceptions of what this type of psychological explanation consists in, but all are agreed that it is strategic and predictive. It is what we use to navigate the social world, just as we use a commonsense physics and a commonsense biology to navigate the physical world. Commonsense psychology is what we use to work out how people will behave in given situations, given what we know of their preferences and the information they have at their disposal. It is what we use to work backwards in explanation from people's behavior to their desires and beliefs, and forwards in prediction from their desires and beliefs to how they will behave. It allows us to work out what people are thinking, to decode their speech, and to integrate our behavior with theirs.

It is frequently suggested, for example, that intelligent behavior can only be explained by appealing to law-like generalizations about the behavior of intelligent agents.[7] These law-like generalizations are formulated in a distinctive cognitive vocabulary and neither they nor the explanations and predictions that they make possible can be captured at lower levels of description. To borrow an example from Zenon Pylyshyn (1981, pp. 4–5), one might appeal in an explanation or prediction to the rule that in the event of an accident one should summon help. One might use this generalization to predict how people will behave if they are first on the scene at a

7 Although, as we shall see in later chapters, this is not the *only* way of understanding how commonsense psychology works. In Chapter 7 we will look at ways of making sense of other people that do not seem to involve this type of law-like generalization – or, for that matter, the conceptual apparatus of commonsense psychology. One of the principal themes of this book is that we should not take standard assumptions about the nature and scope of commonsense psychology for granted. We need to begin, though, by getting some of these standard assumptions clearly in view.

car crash. The generalization appeals to a notion of appealing for help that seems to pick out a clearly understandable set of actions, and we can have some confidence that, whatever particular action is performed by that person, it will fall within the class of actions that can be characterized in commonsense psychological terms as appealing for help. This is important because it looks as if it will not be possible to pick out this class of action in any other way. There is no physical or biological generalization that will pick out all and only the behavioral episodes that might be described as appealing for help and that would count as predictable responses to being the first person on the scene at a car crash. The argument has been made many times (with Putnam 1960; Fodor 1975; Pylyshyn 1984, the best-known examples). It hinges on the idea that a generalization formulated in cognitive terms can be realized in indefinitely many different biological and physical ways on any given occasion. Going for help when one sees an accident can take many different forms. It can involve a series of muscle movements followed by an expulsion of air, or a flailing arm movement directed towards passing traffic, or rotating a dial with a finger or tapping a sequence of buttons. Each of these can in turn be physically realized in indefinitely many ways and nothing links together all the physical descriptions thus generated other than the higher-level fact that they all count as instances of summoning help. Therefore, so the argument goes, it is only at the top level of explanation that this high-level fact can be picked out.

The explanatory level thus identified is the level of commonsense psychology – a form of explanation that is claimed to be predictively adequate and successful on its own terms. Commonsense psychological explanation is thought to be distinctive in two respects:

1 *Distinctive taxonomy*. It involves appeal *at the personal level* to a particular class of cognitive state that does not feature at lower levels in the hierarchy. These are the so-called intentional states – perceptions, beliefs, desires, hopes, fears, and so on. These intentional states play a role in explanation because they have *content* – because they represent the world in certain ways.
2 *Distinctive regularities*. It picks out classes of behavioral regularities that cannot be picked out at other levels of explanation. Typically these will be behavioral regularities only specifiable in commonsense psychological terms – regularities that hold because of the way in which agents represent the world.

Commonsense psychology thus defined is a paradigm of horizontal explanation. There are, as we will see in Chapters 6 and 7, important questions to be asked about the validity of commonsense psychology – some philosophers have argued that our confidence in it is significantly misplaced. And it will emerge that there are good reasons for thinking that commonsense psychology is neither as widely applied nor as widely applicable as it has frequently

been taken to be. But nonetheless, commonsense psychology has a default position at the top of the hierarchy of explanation.

2.4 The interface problem and four pictures of the mind

Now that we have the distinction between horizontal and vertical explanation to hand, and have identified commonsense psychology at the top of the hierarchy of explanation, an obvious question immediately arises. What are the appropriate vertical explanations for the horizontal explanations of commonsense psychology? This is a question about how commonsense psychology interfaces with the levels of explanation lower in the hierarchy. It will be useful to give this question a name. I shall call it the *interface problem*.

> *The interface problem* How does commonsense psychological explanation interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy?

We frequently explain our own behavior and the behavior of those we know and encounter by using the concepts, generalizations and rules of thumb of folk psychology. This has some claim to be the highest level of explanation – the apex of the pyramid. But how does it connect up with the various lower levels of explanation? How do they help to explain it? What vertical connections can we trace downwards from commonsense psychology?

The interface problem is one of the key problems in the philosophy of psychology, and the four pictures of the mind that we shall be exploring in this book can be separated out according to the differing responses they offer to it. Before going on to explore these different responses, it is worth stressing that the interface problem is importantly different from the traditional mind–body problem. The mind–body problem is a metaphysical problem about how mental properties are related to physical properties (or, on an alternative way of putting it, about how mental events are related to physical events), whereas the interface problem is a problem about how (if at all) different levels of explanation relate to one another. It would be perfectly possible for the mind–body problem to be resolved in a way that leaves the interface problem completely *un*resolved. Suppose, for example, that the correct response to the mind–body problem is the view generally known as *token event identity* – the view that each token mental event is identical to some token physical event. This would help us not a jot with the interface problem. Being told that each token mental event is identical to some token physical event does not tell us anything about the connections between the different explanatory projects associated with different ways of looking at that single event. There are, of course, important connections between how one thinks about the ontology of the mind and how one thinks about how to

explain the mind, but the two problems are distinct and can be pursued largely independently of each other.

The four pictures of the mind that we will be discussing in this book offer a spectrum of responses to the interface problem. At one end of the spectrum is what I call the picture of the *autonomous mind*. According to this picture the interface problem is not really a problem at all, since there is a radical discontinuity between explanations given at the *personal level* of commonsense psychology and explanations given at the various *subpersonal* levels of explanation. Subpersonal-level explanations cannot provide a grounding or implementation for personal-level explanations, since there is no equivalent at the subpersonal level of the various constraints of rationality and normativity that govern explanation at the personal level. All autonomy theorists would agree that personal-level explanation only works because of what goes on at the various subpersonal levels (and hence that events at the subpersonal level provide the "enabling conditions" or "conditions of possibility" for personal-level explanation), but they deny that personal-level explanations require *legitimation* or *grounding* at the subpersonal level. The picture of the *autonomous mind* understands the mind in terms of an autonomous and independent type of explanation that has no application to the non-psychological world and that interfaces only indirectly with the types of explanation applicable in the non-psychological realm.

According to the picture of the *functional mind*, however, these differences are exaggerated. Commonsense psychological explanations are a species of causal explanation, no more and no less mysterious than the various types of causal explanation with which we are familiar both from science and from our everyday experience of the physical world. We should understand the intentional states that feature in commonsense psychological explanation in terms of their causal dimension. Mental states have associated with them a determinate causal role, specifying what normally gives rise to them and how they themselves typically give rise to other mental states and to behavior. According to the functional picture of the mind, there are no reasons to think that the interface problem cannot be resolved. Functional approaches to the interface problem adopt one of two strategies. According to the first strategy, which is most popular among philosophers of mind, the network of commonsense generalizations about mental states and behavior that collectively make up commonsense psychology will be matched by an isomorphic network of generalizations holding between physical states. Psychological states are defined by their position in the network of psychological generalizations. They are the nodes of the network. The interface problem is resolved by the existence of systematic relations (relations of *realization* or *implementation*) between the nodes of the psychological network and the physical structures in the brain that serve as the nodes of the isomorphic network at the subpersonal level. According to the second strand of functionalist thinking (what is sometimes called *homuncular functionalism*, but which I will call *psychological functionalism*) solving the interface problem does not require this sort of isomorphism.

Rather, the job of the various subpersonal levels of explanation is to explain the fundamental psychological capacities that are implicated in commonsense psychology. The favored mode of explanation in psychological functionalism is explanation by decomposition, whereby an overarching cognitive task and/or mechanism is broken down into a series of sub-tasks and/or more basic mechanisms, each of which can itself be broken down into further sub-tasks/more basic mechanisms. Different layers of decomposition can be the province of distinct levels of explanation.

The third conception of the mind shares some of the key tenets of the functional picture, but is best considered on its own terms. According to the representational picture, the essence of the mind is indeed given by the causal dimension of mental states, but the interface problem is resolved differently. The key idea behind the representational picture is that psychological states should be understood as relations to sentences in an internal language of thought, where the language of thought is a physically realized medium of thought that has many of the properties of a natural language. The states of commonsense psychology have *semantic properties*. That is, they represent the world in certain ways; they have a certain *representational content*. But these semantic properties are derivative. They are determined by the semantic properties of those "inner sentences". We need to understand a given propositional attitude in terms of the sentence in the language of thought that serves as a surrogate for it in the brain. What gives that propositional attitude its content (what makes it the case that it represents the world in a certain way) is the relation holding between it and objects and properties in the world. This has implications for how we think about thinking. In an obvious sense, thinking involves transitions between psychological states. According to the representational picture, we need to think about thinking in terms of operations that act directly only on the physical properties of those inner sentences, but they do so in a way that preserves sensitivity to the semantic relations between those inner sentences (to the relations that hold between their meanings). The causal transitions between states of the representational mind are purely formal in a way that exactly mirrors the transitions between states of a digital computer. In fact, representationalists effectively claim that the mind can best be modeled as a digital computer.

At the other end of the spectrum from the conception of the autonomous mind lies the picture that I will term the neurocomputational mind. Like representational theorists, proponents of the neurocomputational mind are deeply influenced by the requirements of modeling the mind. They are inspired by a fundamentally different paradigm, however, from representationalists. Whereas the picture of the representational mind is motivated by the idea that the mind is a digital computer and can be studied as a piece of software, in complete independence of the hardware in which it is implemented, neurophilosophers are inspired by research into artificial neural networks. Neural networks are computer models of different types of cognitive ability explicitly designed to reflect certain features of how the brain is

Table 2.2 Key features of the four pictures of mind

	Autonomous mind	*Functional mind*	*Representational mind*	*Neurocomputational mind*
Is there a direct response to the interface problem?	No	Yes	Yes	No
How is the interface problem to be resolved?	—	(1) Different levels of explanation are linked by the relation of *realization* (2) Different levels of explanation operate at different levels of functional decomposition	Psychological states at the personal level are vehicled by sentences in the language of thought at the subpersonal level	Through a dialectic between personal-level explanation and the modeling of personal-level cognitive processes that may well result in the drastic revision of commonsense psychology
Is commonsense psychological explanation to be preserved?	Yes	(1) Yes (2) In a modified form	Yes	Quite possibly not
Is there a genuine hierarchy of explanation?	No	Yes	Yes	Yes
Who is most likely to hold this view?	Philosophers	(1) Philosophers (2) Scientific psychologists	Cognitive scientists, researchers in mainstream Artificial Intelligence	Neuroscientists, researchers in Neural Network Modeling

thought to process information. As we shall see, neural networks do not seem to possess many of the features of commonsense psychological explanations, and this inspires proponents of the neurocomputational mind to stress the discontinuities between personal-level explanation and the neuroscientific explanations occurring at the bottom of the hierarchy. The *de facto* significance in our everyday cognitive life of commonsense psychological explanation is simply a reflection of our ignorance of the real origins and causes of our action – an ignorance that will only be properly addressed at the neuroscientific level. The mind should be modeled as a complex system that may well resist understanding in terms of the crude tools of commonsense psychology.

These four pictures of the mind form a spectrum. The picture of the autonomous mind occupies one extreme and the conception of the neurocomputational mind occupies the other. The centre ground is occupied by the functional picture and the representational picture. Table 2.2 sets out some of the key features of the four pictures. We will be looking at these four pictures of the mind in considerably more detail in the next three chapters – and indeed throughout the rest of the book, for they will be the strands that we will use to explore some of the key issues in the philosophy of psychology.

The next three chapters are largely expository. In Chapter 3 we consider the autonomous mind and the functional mind. I am grouping these together because they present two very different ways of understanding commonsense psychology. In Chapter 4 we turn to the representational mind and we see how proponents of the representational mind think that it emerges naturally as a solution to certain fundamental problems about how the mind can represent the world. The key issue in this chapter is the architecture of cognition – the question of how we should model the subpersonal mechanisms that make cognition and intelligent behavior possible. In Chapter 5 we look at the picture of the neurocomputational mind. I show how it emerges from a rejection of the top-down model of explanation that informs the three other models of the mind. The key tenet of the neurocomputational approach is that personal-level theorizing about the mind and behavior must co-evolve with our understanding of how the brain works. We will investigate the role played by artificial neural networks in carrying forward this co-evolutionary research methodology.

In these three chapters I try to bring out in as much detail as possible the motivations and arguments for each conception of the mind, but evaluation of those motivations and arguments will have to wait until later chapters where we will focus on specific issues and explore the dialectic between these conceptions of the mind as they offer their different approaches to those issues.

3 The nature of commonsense psychology

The autonomous mind and the functional mind

- The autonomous mind and commonsense psychology
- The autonomous mind and the interface problem
- The functional mind
- Philosophical functionalism and psychological functionalism
- Psychological functionalism and the interface problem

Chapter 2 explored the widely held view that the many different levels at which the mind might be studied form a hierarchy, with our commonsense psychological understanding of ourselves and others at the top. But we have not yet gone into much detail about what commonsense psychology actually is. This chapter explores two competing and very different conceptions of commonsense psychology, one associated with the picture of the autonomous mind and the other with the picture of the functional mind. These different conceptions lead to two very different ways of responding to the interface problem.

Section 3.1 outlines the conception of commonsense psychology at the heart of the picture of autonomous mind. The central thesis of the autonomy picture is that there are such radical differences between explanation in commonsense psychology and explanation at lower levels in the hierarchy that there can be no meaningful dialog between the different explanatory projects. As one might expect, this means that autonomy theorists are not in the business of offering direct solutions to the interface problem. Nonetheless, as emerges in section 3.2, the autonomy picture can allow a number of *indirect* responses. Section 3.3 moves on to the functional mind, with particular attention to how its conception of commonsense psychology differs from that at the root of the autonomous mind. Commonsense psychology on the functionalist construal is a causal theory. In section 3.4 we return to the interface problem to see that the general picture of the functional mind can be developed in two different ways, which I term philosophical functionalism and psychological functionalism. Each of these offers a substantively different way of responding to the interface problem.

3.1 The autonomous mind and commonsense psychology

According to the picture of the autonomous mind, commonsense psychological explanations at the top level of the hierarchy are fundamentally different in type and character from explanations lower down in the hierarchy. These differences rule out the possibility of the unified science of the mind that many theorists have envisaged. We can put the point in terms of the distinction between horizontal and vertical explanation developed in section 2.3. According to the picture of the autonomous mind, the vertical explanations holding between commonsense psychological explanations at the personal level and the various different types of explanation operative at the subpersonal level are fundamentally different from the vertical explanatory relations holding between levels of explanation anywhere else in the natural or special sciences.

In order to understand what is distinctive about the picture of the autonomous mind we need to go into more detail about different types of vertical explanation. Recall that vertical explanations aim to provide a legitimation for the horizontal explanations given at a particular level of explanation by grounding them in lower-level explanations. Philosophers of science have closely studied different models of how such vertical explanatory relations might work. One classic type of vertical explanation is what philosophers of science call reduction. Reduction is a relation that holds between theories. In broad terms, the possibility of a reduction exists when one can explain one theory in terms of another. As standardly understood in the philosophy of science, a high-level theory, T1, can be reduced to a low-level theory, T2, when two requirements are met. The first requirement is that there should be some way of connecting up the vocabularies of the two theories so that they become *commensurable* (that is, so that they come out talking about the same things in ways that can be compared and integrated). This is standardly done by means of principles of translation (often called bridging principles) that link the basic terms of the two theories. The second requirement is that the key elements of the structure of T1 should in some sense be derivable from T2, so that T2 can properly be said to explain how T1 works. There are different ways of understanding this second requirement. On the strictest understanding (e.g. Nagel 1961), the derivability requirement is only met when the fundamental laws of T1 (or, more accurately, analogs of the laws of T1 formulated in the vocabulary of T2) can be derived from the laws of T2. When this happens, there is a straightforward sense in which T2, together with the bridging principles, entails T1. A more modest understanding (e.g. Smith 1992) might demand simply that there be an *explanatory interfacing* between the two theories, whereby the reducing theory T2 identifies causal mechanisms that operate to produce patterns identifiable at the level of theory T1.

Proponents of the stronger conception of derivability typically take

examples such as the reduction of the laws of classical thermodynamics to statistical mechanics, while advocates of the weaker construal will more often draw examples from the biological sciences.[1] For example, various parts of the biological sciences employ purposive teleological explanations (appealing, for example, to concepts of design and function) that are manifestly not reducible in any strong sense to causal laws that do not feature teleological concepts. Explanations in biology do not involve laws in anything like the way that explanations in physics involve laws, and hence there is no scope for a strong reduction of the type envisaged by Nagel. But there is nonetheless an interface with non-teleological explanations. An explanation that appeals to the mechanics of natural selection, for example, might explain what makes it appropriate to speak of design and function, while microbiological accounts of the mechanisms of hereditary variation explain how natural selection operates. And so on.

A good way of understanding the core of the conception of the autonomous mind would be as claiming that, irrespective of whether the derivability requirement is understood in strong or weak terms, there are principled reasons for thinking that commonsense psychological explanation is irreducible to any subpersonal level of explanation. Commonsense psychology is radically *incommensurable* with all subpersonal theories, in virtue of employing a distinctive type of explanation that cannot in any way be integrated with the types of explanation operative at the subpersonal level. Let me give a brief characterization from one of the leading contemporary autonomy theorists of what the contrast is supposed to consist in, before going on to explain the alleged incommensurability in more detail. Here is John McDowell:

> The concepts of the propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be. This is to be contrasted with a style of explanation in which one makes things intelligible by representing their coming into being as a particular instance of how things generally tend to happen.
>
> (1985, p. 389)

It is easier to get the general flavor of what is going on here than to spell it out in any detail. But the basic idea is that commonsense psychology is autonomous because, unlike the levels of explanation lower in the hierarchy, it is essentially *hermeneutic*. It explains intelligent behavior by interpreting it as the behavior of rational agents. The principles of rationality regulating the interpretation of rational agents are normative principles rather than descriptive generalizations (principles that describe how people ought to

1 In fact, as we will see in section 5.1, even the classic example of thermodynamics and statistical mechanics is far less straightforward than many authors have suggested.

behave, as opposed to descriptions of how they generally do behave). In so far as we take a piece of intelligent behavior to be the behavior of a rational agent, we try to make sense of it so that it comes out as the rational, appropriate and comprehensible thing for the agent to do, given what we know of what she wants to achieve and of what she believes about the world. But in determining what the agent wants to achieve and what she believes, we also need to interpret her speech and behavior so as to attribute to her a consistent and largely truthful set of beliefs together with a coherent and realistic set of desires and preferences. The process of interpretation is essentially a process of rational reconstruction aiming to maximize the rationality of the agent whose behavior is being interpreted (Davidson 1970, 1974; Putnam 1983).

The autonomy of personal-level commonsense psychology comes in because there is no analog of these general norms of rationality, consistency and coherence in any of the forms of explanation operating "below the level of the person" – at what we described in the previous chapter (section 2.2) as the subpersonal level of explanation. Donald Davidson, another prominent autonomy theorist, draws the contrast as follows:

> Any effort at increasing the accuracy and power of a theory of behavior forces us to bring more and more of the whole system of the agent's beliefs and motives directly into account. But in inferring these systems from the evidence, we necessarily impose conditions of coherence, rationality and consistency. These conditions have no echo in physical theory.
>
> (Davidson 1974, p. 231, in Davidson 1980a)

In contrast to explanation at the personal level, the various different types of subpersonal explanation, from neurobiology to computational systems theory, are descriptive rather than normative. Their concern is with subsuming particular events under general laws – with, as McDowell puts it, "making things intelligible by representing their coming into being as a particular instance of how things generally tend to happen".

An example will make the point more vivid. An experimental psychologist interested in developing a model of decision-making, for example, will be interested primarily in capturing experimentally detectable regularities in how people actually go about the business of practical reasoning. So, for example, it is well documented (as we will see in more detail in section 8.4) that people regularly make various fallacious inferences in both deductive and probabilistic reasoning (Evans and Over 1996). The experimental psychologist whose concern is modeling people's actual decision-making will of course want to build these into departures from the norms of rationality into his theory. Were he not to do so, his model would be descriptively and empirically inaccurate. From the viewpoint of commonsense psychology (understood as the autonomy theory understands it), in contrast, we view agents as rational beings. We attempt to make sense of their behavior, to

understand what they will do and why they did what they did, on the assumption that they are rational agents. Without that assumption we will have no way of getting from what we know of their preferences and beliefs to their behavior. But, we can only make predictive and explanatory use of that assumption if we abstract away from the *descriptive* details of how they might actually be reasoning on a particular occasion and attend instead to the *prescriptive* or *normative* dimension of how they ought to reason.

The autonomy theorist's central claim, therefore, is a combination of two basic theses. First, there is an irreducibly normative dimension to commonsense psychological explanation, as a function of the constitutive role played by ideals of rationality, consistency, and so forth. Second, there is "no echo" of this normative dimension at any subpersonal level of explanation. All autonomy theorists are agreed that the combination of these two theses entails a radical incommensurability between commonsense psychological explanation, on the one hand, and anything that might be described as scientific psychology (in the broad sense sketched out in Chapter 1). However, there are two rather different ways of developing the picture of the autonomous mind and, correspondingly, two different ways of understanding the interface between commonsense psychology and scientific psychology. These will be the subject of the next section.

3.2 The autonomous mind and the interface problem

Recall that the interface problem is the problem of explaining how the horizontal explanations of commonsense psychology interact vertically with levels of explanation lower down in the hierarchy of explanation. It is clear that most of the standard ways of responding to the interface problem are unavailable to autonomy theorists. The radical incommensurability between commonsense psychology and the various subpersonal levels of explanation is incompatible with responding to the interface problem in the manner either of weak or of strong reduction.

A strong reduction is clearly ruled out. A strong reduction is only available where the central principles of the theory to be reduced (i.e. commonsense psychology) can be formulated employing the concepts and laws of the reducing theory. But the autonomy theorist cannot allow that this could ever be possible, whatever candidate theory is selected from the domain of the subpersonal, given that the constitutive norms of rationality governing commonsense psychology are unavailable at the subpersonal level. Since subpersonal-level theories trade in the descriptive rather than the normative, there is no way that they could possibly have the resources to capture norm-laden and rationality-governed explanations at the personal level.

But nor will a weak reduction be available to autonomy theorists. Recall that a weak reduction does not seek a reduction of the basic principles of one theory to the basic principles of another theory. Rather, weak reductions aim

to identify at the lower level the mechanisms responsible for the emergence of the patterns discernible at the higher level.[2] Autonomy theorists cannot allow that there are such explanations, however. The patterns discernible at the level of commonsense psychology are not patterns produced by a mechanism of the sort that might be explained in, for example, computational terms. In fact, they are not patterns produced by a mechanism at all. Rather, they are abstract patterns that emerge when we think about the demands and requirements of reason, consistency and coherence, and about how the corresponding norms might be applied in particular situations. So, to return to our earlier example, autonomy theorists need have no quarrel with the suggestion that there may be a practical decision-making module in the brain, responsible (say) for computing how best to maximize expected utility in a given situation. What they would stress, however, is that, whether or not such a mechanism exists, it has no role to play in explaining the patterns of rational, coherent and consistent behavior that we (as commonsense psychological explainers) identify at the personal level – since these patterns are not patterns in how people *actually* go about reasoning and making up their minds, but rather in how they *ought* to do so.

What I have characterized as weak and strong reductions are not the only ways of responding to the interface problem. We will be looking at other proposals later on in this chapter. But it should already be clear that the radical incommensurability between the conceptual framework of commonsense psychological explanation and that of levels of explanation lower down in the hierarchy is a serious obstacle to any head-on response to the interface problem. Unsurprisingly, autonomy theorists have not given an enormous amount of thought to the interface problem – from their theoretical perspective it is not a particularly pressing problem. Nonetheless, within the conception of the autonomous mind it is possible to identify two different ways in which autonomy theorists have reacted to the interface problem. One of these responses effectively denies that there is any interface at all between commonsense psychology and any subpersonal realm of explanation. This position has been most comprehensively worked out in the writings of John McDowell and Jennifer Hornsby. Donald Davidson's much-discussed doctrine of anomalous monism, however, provides autonomy theorists with the resources for a less drastic response to the interface problem.

Let us start with the less drastic response. Davidson's doctrine of anomalous monism is the best-known development of the picture of the autonomous mind (Davidson 1980a). The concerns lying behind the theory have to do primarily with the causal dimension of commonsense psychological explanation. Commonsense psychological explanations work by citing particular beliefs and desires that jointly render it comprehensible (i.e.

2 This conception of abstract patterns holding at the personal level of explanation is particularly associated with Daniel Dennett and will be discussed further in section 6.1.

rational from the agent's point of view) why an agent performed a particular action – just as commonsense psychological predictions work by offering a particular course of action as rational in the light of the agent's desires and the information available to him. But, Davidson stresses, what makes these genuine explanations (as opposed to mere rationalizations after the event) is that they identify the beliefs and desires that actually caused the agent to behave in the way that he did. Commonsense psychological explanation (for Davidson and for many others) is a species of causal explanation (Davidson 1969). Yet this poses an immediate problem. According to influential models of causal explanation (including Davidson's own), such explanation depends crucially upon the existence of causal laws (Davidson 1970). We can only say that an event of a particular type causes an event of another particular type if there is a law to the effect that events of the first type are always followed by events of the second type. But where are we to find these laws in the domain of commonsense psychological explanation? The problem is particularly acute for defenders of the picture of the autonomous mind, since their principal claim is that the realm of commonsense psychology is not governed by descriptive *laws* at all. The normative relations of rationality, coherence and consistency holding between intentional states in commonsense psychological explanation are not at all law-like in the manner required to underwrite causal laws. They do not, as we have seen, describe how things are. Nor are they exceptionless. They hold only "for the most part", subject to numerous and uncodifiable exceptions. They are generalizations, but not laws.

The problem, then, is as follows. On the one hand, commonsense psychological explanation is supposed to be causal, and hence governed by causal laws. On the other hand, the normative principles that govern such explanation are anything but causal and in fact seem to rule out the possibility of such causal laws. Davidson's solution to the problem is ingenious. He rejects the way the problem is set up. We cannot, strictly speaking, talk of physical events and psychological events, with one class of events but not the other featuring in strict causal laws. Causation is a relation that holds between events simpliciter, whereas causal laws hold over events only when they are described in particular ways. One and the same event can be described in both physical and psychological terms, and might fall under a law under one description but not another. So, Davidson argues, even though the psychological events featuring in psychological explanations cannot feature in causal laws *under the description in which they feature in commonsense psychological explanations*, they are nonetheless identical to physical events that do cause the behavior being explained. When characterized in physical terms, these events do indeed fall under strict causal laws, thus vindicating the causal explanatory relations in which, under their psychological descriptions, they stand to the behavior being explained. The generalizations of commonsense psychology are not themselves law-like, because they are irreducibly qualified and imprecise. There is no prospect of them being made precise without shifting out of the open-ended vocabulary of the propositional attitudes and

into the closed vocabulary of physics.[3] Nonetheless, they lend support to commonsense psychological explanations by pointing to the existence of genuine law-like regularities – albeit ones that can only be identified by shifting from psychological vocabulary to physical vocabulary.

The interface problem is thus solved by postulating token-identities between intentional states and physical states that allow commonsense explanations citing those intentional states to be genuine causal explanations.[4] It is important that the postulated identities are token identities (holding between individual intentional states and individual physical states) rather than type-identities (holding between types of intentional state and types of physical state). If the identities were type-identities, then that would create precisely the sort of systematic law-like vertical correlations between personal-level states and subpersonal-level states that the autonomy theorist denies are possible. This would, moreover, open up the possibility of law-like horizontal relations holding within the realm of commonsense psychology – since these would be direct consequences of the law-like horizontal relations holding between subpersonal-level states. In contrast, the thesis of token-identity allows the personal and subpersonal levels of explanation to interface in virtue of explaining the same event, even though they do so under radically different descriptions and with dramatically different resources.

A simplified example will illustrate how anomalous monism is supposed to work. Let us imagine that one belief causes another – say, that my belief that Edinburgh is north of Paris causes me to have the further belief that Edinburgh is north of Berlin. For reasons that we will be exploring, Davidson holds that there can be no strict causal law to the effect that believing that Edinburgh is north of Paris will cause one to believe that Edinburgh is north of Berlin. Nonetheless, each belief is (let us assume) identical to some neurophysiological state. So, the causal connection between the two mental events just is the causal connection between the two neurophysiological events. There is no obstacle, Davidson thinks, to there being a causal law connecting types of physical event, *provided that those physical events are characterized in physical terms*, and there are, Davidson thinks, causal laws defined over neurophysiological events. These causal laws, holding over events that are in fact mental events even though they are not characterized in mental terms, allow the causation of one belief by another to satisfy the principle of the nomological character of causation. A similar, although somewhat more complex, account will hold for cases where a combination of mental states causes behavior.

An objection frequently leveled at Davidson's account is that it fails properly to explain mental causation, because it does not allow mental events to be causally efficacious in virtue of their mental properties (Honderich 1982). It is not, for example, the fact that my belief is a belief about the geographical

3 Davidson's arguments for this claim (for the so-called *anomalism of the mental*) will be discussed in section 6.2.

4 For further discussion of the different types of identity theory, see Kim (1996, Chapter 3).

location of Edinburgh relative to Paris that causes me to have the further belief that Edinburgh is north of Berlin. The causal connection holds, rather, in virtue of the firings of neurons and the activities of neurotransmitters. To put it in terms that we will be employing in Chapter 4, anomalous monism does not seem able to accommodate *causation by content*. This objection, although frequently taken to be devastating to Davidson's position, is in fact rather question begging. Anomalous monism is formulated in the context of a theory of causation and events on which it does not make sense to talk of one event causing another in virtue of its properties. Causation is a *metaphysical* relation holding between events. Events themselves do not, on Davidson's view, have properties. The properties of events only come into the picture when those events are characterized in certain ways. Properties are relevant in the context of explanation, but not, Davidson thinks, in the context of causation. Of course, Davidson's way of thinking about events can be, and has been, criticized, but its falsity can hardly be presupposed in a criticism of anomalous monism.

A better objection to anomalous monism highlights the role of causal laws in explanation and prediction. One of the hallmarks of a genuine causal explanation is that it supports counterfactuals. If *F* causes *G*, then it is natural to conclude that, had there not been an *F*, a *G* would not have occurred. It is natural to think, and many philosophers have thought, that it is the fact that causal explanations are governed by causal laws that explains why the counterfactuals hold. If there is a causal law to the effect that *F*-type events cause *G*-type events, then we can assume, provided the appropriate background conditions hold, that if an *F*-type event *were* to take place, it *would* cause a *G*-type event – and that, were there not to be an *F*-type event, there would not be a *G*-type event. This poses a problem for Davidson's theory, however. The central feature of anomalous monism is that the causal explanation and the supporting causal law are formulated in completely different terms. The causal law is, we have assumed, a law spelling out the relation between neurophysiological states, while the causal explanation is formulated in the language of propositional attitude psychology. The causal explanation does, of course, support counterfactuals, which are themselves formulated in the language of propositional attitude psychology. Suppose we ask what explains those counterfactuals. It is hard to see why counterfactuals about what would happen were someone not to believe, for example, that Edinburgh is north of Paris should be underwritten by a causal law governing the relation between types of neurophysiological event. The causal law can tell us only about what would happen if one type of neurophysiological event were not to occur. And it is important to realize that Davidson's theory rules out an intuitively appealing solution to this problem. Davidson can identify individual mental events with individual neurophysiological events, but it is not open to him to identify types of mental event with types of physical event (so that counterfactuals holding over neurophysiological events would *ipso facto* be counterfactuals holding over belief states). David-

son's theoretical commitments allow him to be a token-identity theorist, but not a type-identity theorist. The reason for this is straightforward. Type-identities would make possible precisely the sort of strict laws connecting the physical and the psychological that are ruled out by the anomalism of the mental. It would seem, therefore, that Davidson is caught on the horns of a dilemma. Either his causal laws fail to support the right sort of counterfactuals, or his account of mental causation comes into conflict with the anomalism of the mental.

Even putting these difficulties to one side, Davidson is clearly offering a metaphysical way of resolving the interface problem. The problem is tackled by suggesting that there is a metaphysical relation (namely, identity) between the *explananda* of the different levels of explanation. As such, however, it hardly does justice to the original thought behind the interface problem. Telling us how the *explananda* of different theories and levels of explanation are related to each other does not tell us anything about how those explanations themselves mesh together – and hence has nothing to contribute to the project of constructing a unified picture of the mind that draws together the many different levels at which it can be studied. There are many reasons why one might be dissatisfied with anomalous monism as a metaphysical thesis – and in particular as a solution to the problem of how commonsense psychological explanations can be causal explanations. Critics of Davidson have frequently suggested, for example, that his theory makes mental states *epiphenomenal* (see, for example, the essays in Heil and Mele 1993). That is to say, the properties of mental states that are causally effective are not those features that have a role to play in commonsense psychological explanation. Anomalous monism makes intentional states causally effective, but not *qua* intentional states. Even if we ignore all these criticisms, however, anomalous monism does not really address the concerns that give rise to the interface problem. In fact, its stress on the incommensurability of the personal and subpersonal levels seems to entail that little, if anything, can be said about the relation between the different levels of explanation.

The second way of developing the autonomy theorist's basic thesis is even more uncompromising in its approach to the interface problem. John McDowell and Jennifer Hornsby have worked out a conception of the distinctiveness of commonsense psychological explanation that rejects even the weak thesis of token-identity (Hornsby 1980–81, 1986; McDowell 1985). On their view there is no connection whatsoever between the respective *explananda* of commonsense psychology, on the one hand, and the various levels of subpersonal psychology (broadly construed), on the other. Commonsense psychological explanation is, quite literally, talking about different things and events from the various subpersonal levels of explanation. Here is how Hornsby characterizes the position:

> Subpersonal accounts which are introduced to explain the results of investigations are not to be thought of as providing a new understanding of

> that which commonsense psychology previously explained. The questions that laboratory psychologists answer when they do the kinds of experiments that lead to subpersonal theories are not the questions that we can know the answers to by interacting, as commonsense psychological subjects, with others ... It is because the everyday Why-questions which are answered using commonsense psychology require one to operate with a conception of a subject as rationally motivated as one is oneself that the accounts of subpersonal psychology must be addressed to a different set of *explananda*.
>
> (1997, p. 167)

Of course, to say that commonsense psychology and subpersonal psychology have different *explananda* is not to make any concessions to dualism. There is no suggestion that the types of states and events in which commonsense psychology deals are in any way non-physical. The thought is rather that, although they are physical states and events, they do not map in any systematic way onto states and events that might be studied at lower levels of explanation.

The position here can be understood as a development of certain aspects of Davidson's arguments for anomalous monism. Theorists such as McDowell and Hornsby accept Davidson's arguments for the anomalousness of the mental. In particular, they stress the irreducibility of the normative dimension of our personal-level psychological concepts and explanations. But they extend these arguments in a way that rules out the possibility of personal-level psychological states being token-identical with subpersonal states. The issue is one about how to individuate personal-level states. Davidson is quite happy to say that there is a single event that can be characterized either physically or psychologically. Hornsby and McDowell, on the other hand, think that the very same considerations that point to the irreducibility of the psychological also show that mental events must be individuated in fundamentally different ways from physical events – and hence that token-identity is not a live option. The only physical events that could be identified with mental events are complex neurophysiological events, and we cannot view the rationality of an agent's action in the light of the agent's beliefs and desires in terms of the relation between a set of bodily movements and the set of neurophysiological events that generates those bodily movements. The rationality of an action is a function of how the person as a whole behaves – not of the causal ancestry of a set of bodily movements (even were one able to identify that causal ancestry). Nor are theorists of this stamp moved by Davidson's claim that to deny token-identity is effectively to deny the causal efficacy of the mental. As we will see in the next section, this version of the autonomy picture goes naturally with a much less demanding way of understanding mental causation – the counterfactual approach to mental causation, which rejects the idea that genuine causation requires the existence of causal laws (the so-called nomological character of causation).

As far as the interface problem is concerned, it is clear that autonomy theorists of this type hold that the level of commonsense psychological explanation does not require vindication or legitimation from lower levels of explanation – even vindication of the minimal type provided by the thesis of token-identity. It is unsurprising, therefore, that they deny that there are any vertical explanatory relations holding between the top level of the hierarchy and the lower levels. Nonetheless, and this is the distinctive twist, the top level does not float completely free. Although there are no vertical explanatory relations holding between the levels, there remains a degree of *explanatory relevance* holding between horizontal explanations at the top level and horizontal explanations at lower levels. Horizontal explanations at the lower levels explain the *enabling conditions* of commonsense psychological explanation. As Hornsby puts it, "a subpersonal account shows how it can be that something has the various capacities without which nothing could be the sort of commonsense psychological subject that a person is" (1997, p. 166).

This is somewhat vague, and no autonomy theorist has given a positive account of the notion of an enabling condition. But this is what one would expect, given that the autonomy theory's main concern is with trying to persuade philosophers and psychologists that the interface problem is far less pressing than it might immediately appear. The autonomy theory is primarily a negative theory. In fact, the main negative claim that it is trying to put across is effectively that there can be no such thing as the philosophy of psychology in the sense that I have characterized it in Chapter 1. I suggested there that the philosophy of psychology is essentially the interdisciplinary enterprise of developing a unified account of the central concepts that feature both in our commonsense conceptual scheme and in the scientific study of cognition. If the autonomy theory is taken at face value, however, there are no such concepts. The conceptual scheme of commonsense psychology is completely insulated from the various conceptual schemes implicated in the scientific study of cognition. There are no concepts that feature both in commonsense psychology and in any of the subpersonal levels of the hierarchy of explanation.

The crucial issue in evaluating the autonomy theory is how plausibly it characterizes commonsense psychology. Is commonsense psychological explanation really as incommensurable with the types of explanation offered in scientific psychology as autonomy theorists claim? We will be discussing the nature and significance of commonsense psychology in Chapters 6 and 7. The discussion there will put considerable pressure on the autonomy theory. In particular, it will be suggested both that the domain of commonsense psychology is far more circumscribed than it is taken to be by autonomy theorists and that those theorists greatly overplay the contrast between the normative dimension of commonsense psychology and the descriptive nature of subpersonal explanation. For the moment, however, here is a reminder of some of the key points about the autonomous picture of the mind.

Checklist for the autonomous mind

- The key tenet of the autonomous conception of the mind is that there is a radical incommensurability between the type of explanation at play in commonsense psychology and that involved in explanation at the subpersonal level.
- This incommensurability is claimed to derive from the centrality in commonsense psychological explanations of the normative ideals of rationality, coherence and consistency. We explain why people behave as they do (and predict how they are going to behave) on the assumption that they are rational agents with coherent and consistent sets of beliefs and desires.
- According to autonomy theorists, there is nothing at the subpersonal level that can capture the role of these normative ideals of rationality, consistency and coherence.
- Davidson's anomalous monism is one way of developing the autonomy theory. It maintains that, although the modes of explanation operative at the different levels of explanation are different and incommensurable, the different levels of explanation are still explaining the same thing under different descriptions.
- A more extreme version of the autonomy theory, associated with John McDowell and Jennifer Hornsby, denies the claim of token-identity characteristic of anomalous monism. The *explananda* of commonsense psychology do not feature in any way at all at the subpersonal level.

3.3 The functional mind

As with the autonomy conception, the picture of the functional mind accords a privileged role to commonsense psychological explanation in the understanding of the mind. Functionally-minded philosophers and psychologists place great emphasis on the claim that commonsense psychological explanations allow us to detect patterns of behavior that are simply invisible to levels of explanation lower down in the hierarchy of explanation (although they characterize these patterns very differently). Functionalists differ fundamentally from autonomy theorists, however, in two related respects. The first is that they do not make such a sharp distinction between the generalizations of commonsense psychology and "ordinary" causal generalizations. Without denying that commonsense psychology assumes the rationality of the agents whose behavior it is trying to explain or predict, the functional picture of the mind denies that this makes commonsense psychological explanation qualitatively different from explanation at the subpersonal level. The generalizations of commonsense psychology are not different in kind from generalizations lower down in the hierarchy of explanation. Correspondingly (and this is the second difference) advocates of the functional picture have the resources to respond directly to the interface

problem. On the functionalist picture the interface problem is resolved in one of two ways, depending on which strand of functionalism is in play. Before looking at how the interface problem is resolved for the functional mind, however, we need a firm grip on how commonsense psychological explanation is understood on the functional picture.

The most fundamental difference between the autonomous mind and the functional mind has to do with the causal dimension of the mind. The issue is not whether psychological explanation is causal explanation. It was for a time fashionable in the 1950s and 1960s (particularly among philosophers inspired by Wittgenstein) to argue that psychological explanations looked for the reasons for which agents performed actions, rather than the causes of those actions, but few philosophers would nowadays deny that reason-giving explanation is a species of causal explanation. The point at issue is how the causal dimension of commonsense psychological explanations is to be understood. In the previous section we have already seen one way in which an autonomy theorist might attempt to cash out this causal dimension. Donald Davidson's anomalous monism offers an account of how personal-level explanations can qualify as causal. According to anomalous monism, the psychological states invoked in personal-level explanations are token-identical to physical structures that themselves stand in law-governed causal relations to the action being explained (or predicted).

This is not the only way the autonomy theorist can allow for causation at the personal level. Another available strategy would be to challenge the common conception of causal explanation as involving the subsumption of individual events under causal laws. It will be remembered that this conception of the so-called *nomological* nature of causation (from νομος, the Greek word for a law) is what makes the idea of causation at the personal level so problematic for the autonomy theorist – because the autonomy theorist denies that there are any causal laws holding at the personal level. The autonomy theorist might accordingly offer a non-nomological account of causation. The only candidate theory here that has been worked out in any detail is the *counterfactual* account of causation (Hornsby 1997; Baker 1995). The basic idea is that a particular combination of mental states causally explains a given behavior if and only if it is true that in the absence of that combination of mental states the behavior in question would not have occurred – and, moreover, that that same combination of mental states would have led to the behavior in question even in different circumstances and background conditions. This is called the counterfactual theory because it makes the existence of causal relations dependent upon the truth of conditional statements that are counterfactual (that is, statements about what *would have* happened *if* the starting conditions had been different from how they actually are).

As we saw in the previous section, there are genuine questions to be asked about whether anomalous monism really captures the causal dimension of psychological explanations, since the causal weight does not seem to be

borne by mental states *qua* mental states. And there are many potential reasons for dissatisfaction with the counterfactual approach. The most fundamental difficulty is that it seems to get the order of explanation the wrong way round. It is certainly true that the existence of a genuine causal relation is closely linked to the truth of certain counterfactuals. It cannot be the case that event *E* causes event *E** unless it is true that, had there not been an *E*-type event there would not have been an *E**-type event.[5] But this counterfactual dependence of effect on cause does not exhaust the nature of the causal connection between *E* and *E** – rather, it is itself explained by the fact that event *E* has caused event *E**. It is natural to think that the counterfactual dependence holds because there is causation, rather than vice versa.[6]

We will return to the causal dimension of commonsense psychological explanation in the next chapter. For the moment the important point is that the position in logical space occupied by the functionalist should be clear. Like the anomalous monist, but unlike the counterfactual theorist, the functionalist holds that a genuine causal explanation requires the existence of a causal law connecting the relevant two events. Unlike the anomalous monist, however, he holds that this causal law must be a causal law holding at the level of commonsense psychology – a law connecting different types of propositional attitude; connecting a certain type of input with a certain type of propositional attitude; or connecting a particular type of propositional attitude with a particular type of behavior. The generalizations of commonsense psychology are, quite simply, causal generalizations and the explanations and predictions offered by commonsense psychology are causal explanations that should be understood in the way that causal explanations have classically been understood – namely, as involving the subsumption of two events under a general causal law.

Figure 3.1 presents the different theoretical possibilities here. As the argument-tree makes clear, one arrives swiftly at functionalism if one thinks (a) that commonsense psychological explanations are causal explanations; (b) that causal explanations require the existence of causal laws; and (c) that the causal laws governing commonsense psychological explanation have to hold at the personal level.

Of course, the functionalist still has to deal with the reasons philosophers have given for denying that there could be causal laws operating at the personal level. These reasons fall into two broad groups. The first group cluster around the general idea (introduced in the context of the autonomy theory in the previous section) that commonsense psychological explanation is not a

5 This is not strictly true. The occurrence of the *E**-type event may have been *pre-empted* or *overdetermined* – that it is to say, preceded or accompanied by a second event such that, had the *E*-type event not occurred, the second event would still have been sufficient to bring about the *E**-type event. Counterfactual theories of causation typically have difficulty accommodating pre-emption and overdetermination.

6 The counterfactual theory is discussed in more detail in section 6.3.

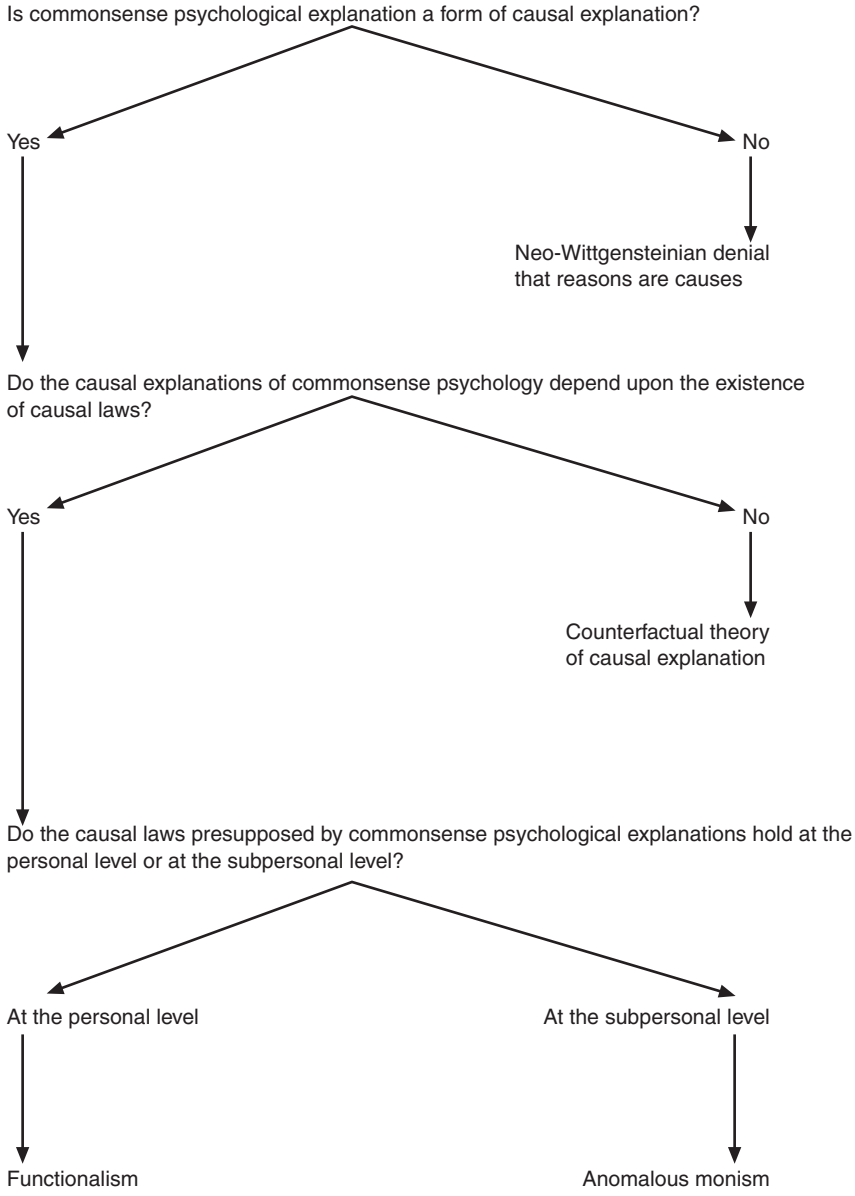

Figure 3.1 Psychological explanation and causation: theoretical possibilities.

descriptive enterprise of the sort that might feature causal generalizations, but rather a normative enterprise of interpretation in which descriptive generalizations have no place. The second group of reasons emphasize the lack of robustness of the most obvious candidates for the status of causal generalizations of commonsense psychology. These generalizations are (it is claimed) more like rules of thumb than full-blooded causal laws. They do not hold universally and it is impossible fully to specify the circumstances in which they not hold (in the way that many have thought it is always possible to do with genuine scientific causal laws). In effect, all one can say is that they are generalizations holding *ceteris paribus* (all other things being equal) and it is frequently suggested that no such generalizations can be genuinely law-like (Schiffer 1991).

I shall postpone to later chapters detailed discussion of how defenders of the idea that the generalizations of commonsense psychology are causal generalizations might respond to these lines of argument.[7] But in very broad outline one might expect to see the following lines of reply. In response to the charge that the generalizations of commonsense psychology are normative rather than descriptive, the defender of the causal thesis is likely to reply that the line between normative and descriptive is far less clear than the autonomy theorist assumes. In particular, it might be suggested that commonsense psychological explanation and prediction could not possibly work as well as they do were they not sensitive to the basic descriptive facts about how people tend to behave in particular circumstances. There must be more going on in psychological explanation than simply reading off from the norms of rationality how one might expect a perfectly rational agent to behave. Commonsense psychological explanations have both a normative and a descriptive dimension. Developing this line of thought meshes naturally with a response to the second cluster of objections. It would be a mistake (someone might suggest) to draw too sharp a contrast between the *ceteris paribus* generalizations of commonsense psychology and scientific causal laws. Even the laws of physics are *ceteris paribus* laws, since they are formulated for idealized situations that never arise even in the laboratory, let alone in the real world (Cartwright 1983; Huttemann 2004).

Whatever the ultimate outcome of the debate, it is clear how things stand according to the picture of the functional mind. The claim that the generalizations of commonsense psychology are causal generalizations eliminates the alleged differences between explanations at the personal and at the subpersonal level and restores the *commensurability* between personal and subpersonal levels denied by autonomy theorists. Equally importantly, it permits functionalists to develop their distinctive characterization of mental states. The key idea of functionalism is that mental states are defined in terms of how they feature in psychological causal laws. Consider, for example, the mental state of having a headache. This state will feature in a range of causal

7 These topics are discussed in some detail in Chapter 6, particularly section 6.2.

laws specifying the typical causes and effects of being in a headache (both its effects on behavior and its effects within the cognitive economy). These causal laws collectively define the functional role (or: the causal role) of being in a headache. The same holds, according to functionalists, for all mental states. Each mental state has associated with it a functional/causal role given by the causal laws that specify the typical causes and effects of that state. Many of the causal laws governing psychological explanation will spell out the causal relations between different mental states – how one mental state will typically give rise to another, for example. So, the functional roles of different mental states will typically be interdependent.[8] The picture that emerges is of commonsense psychology forming a theory that defines a set of functional roles. The functional roles associated with different mental states are the nodes in the network of causal generalizations yielded by commonsense psychology.

One significant benefit of thinking about mental states in terms of functional roles is that functional roles are *multiply realizable*. A range of completely different physical structures can realize the same functional role by performing the basic functions that define that role. Anything can occupy a given functional role, provided that it stands in the appropriate set of causal relations fixed by the causal laws that determine the functional role. What makes this possible is that functional roles are characterized in terms that abstract away from the physical details of how they might be implemented. This is very important, since all the evidence is that certain mental states and their corresponding functional roles are realized differently in humans and other species – the human perceptual systems, for example, are fundamentally different from many to be found in the animal kingdom. And it may well be the case, given what is known about the plasticity of the brain, that even within the human population there are variations in realizers for certain mental states – due to brain damage, for example, or simply as a function of different stages in development.

Moreover, and this is the key to how one important strand within the

8 This gives rise to concerns about circularity very similar to those that bedevilled philosophical behaviorism. How can one define one mental state in terms of another when the second mental state is itself defined in terms of the first? Functionalist philosophers, pursuing a suggestion made in a different context by Frank Ramsey, have developed a technical way of dealing with this difficulty – the technique of ramsification. The basic idea of ramsification is to give a comprehensive theoretical statement of commonsense psychology where the names of mental states are each replaced by a corresponding variable. The variable that stands in for the name of a given belief will stand in for it in every law in which it features. We end up, therefore, with a statement of schematic theory that contains only variables. Each variable is implicitly defined by its role in the theory. The next step is to bind these variables by existential quantifiers to yield a theory effectively stating that each theoretical role is uniquely satisfied. When the theory of commonsense psychology is formulated in these terms individual mental states will effectively be defined in terms of each other without circularity. The method of ramsification was first applied to commonsense psychology by David Lewis (1972). Further details will be found in textbooks on the philosophy of mind, such as Kim (1996) and Rey (1997).

functional approach to the mind proposes to tackle the interface problem, specifying functional roles offers a way of bridging the gap between personal and subpersonal levels of explanation. A functional role is an abstract specification of a particular set of causal relations. These causal relations hold at the subpersonal level and we can use this fact to identify subpersonal realizers for personal-level functional roles. Part of the functional role of having a headache, for example, is that it should have certain typical causes (a sharp blow to the head, for example, or dehydration) and certain typical effects (such as leading the sufferer to avoid bright lights and loud noises). We can, so the theory goes, identify the realizer of that functional role (within a given population) by identifying the subpersonal state that has those typical causes and typical effects – that is to say, the neural state that is typically caused by, among other things, sharp blows to the head and typically causes, among other things, noise-avoiding behavior. More generally, we can identify at the subpersonal level a network of causal generalizations defined over subpersonal states that is isomorphic to the network of causal generalizations at the personal level defined by commonsense psychology. The nodes in the subpersonal-level network are the realizers of the roles fixed by the nodes in the personal-level network. The personal level is the level of roles – of abstract specifications of causal relations. The subpersonal level, in contrast, is the level of realizers – where the causal work identified at the personal level actually gets done. So, a proper understanding of commonsense psychological explanation gives us a blueprint for making sense of what is going on at the subpersonal level.

We need, however, to distinguish two different ways in which this general strategy can be used to respond to the interface problem within the general picture of the functional mind, corresponding to two broadly different ways in which the functional picture can be developed. I will call these *philosophical functionalism* and *psychological functionalism* respectively. These two different types of functionalism vary in two dimensions. The first dimension is how they go about identifying the causal generalizations of commonsense psychology, while the second concerns their respective understanding of the vertical relations holding between those causal generalizations and the various subpersonal levels of explanation. We will look at these two ways of responding to the interface problem in the next section.

3.4 Philosophical functionalism and psychological functionalism

Many proponents of philosophical functionalism hold that the causal generalizations of commonsense psychology can effectively be read off from our everyday understanding of ourselves. On this view, commonsense psychology is (at least in principle) fully transparent to ordinary psychological subjects. In so far as we are normally functioning psychological subjects (as opposed, say, to being autistic or simply too young to have developed the

necessary skills for navigating the social world), we all have an implicit grasp of the fundamental principles of commonsense psychology. An adequate formulation of commonsense psychology will emerge once we make those fundamental principles explicit. One way of understanding this general claim comes across very vividly in the following passage from David Lewis:

> Think of commonsense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli and motor responses ... Add also all the platitudes to the effect that one mental state falls under another – "toothache is a type of pain" and the like ... Include only platitudes which are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that the names of mental states derive their meaning from these platitudes.
>
> (1972, p. 212)

Commonsense psychology is a theory that we all share and whose veracity we can all affirm when it is made suitably explicit. And commonsense psychology is equally a guide to the nature of mental states. We can use the functional roles fixed by the nodes of the commonsense network of causal generalizations to identify the realizers of the mental states – and hence to bridge the gap between personal and subpersonal levels of explanation.

This type of philosophical functionalism (sometimes called *folk functionalism* – e.g. by Rey 1997, Chapter 7) contrasts with *a priori* or *conceptual* functionalism, which holds that the causal generalizations of commonsense psychology can be derived from our everyday psychological concepts by *a priori* conceptual analysis (see the essays in Shoemaker 1984). The difference is not simply one of emphasis. Both types of functionalism tie the meanings of our everyday psychological vocabulary to the role that the corresponding concepts play in commonsense psychology, but folk functionalism does not take the principles of commonsense psychology to be *a priori*. Folk functionalism, unlike *a priori* functionalism, takes the concepts of commonsense psychology to be law-cluster concepts (in roughly the sense outlined in Chapter 1, although with the crucial difference that the laws in question are taken only from the personal level of explanation). Nonetheless, both varieties of philosophical functionalism share the basic assumption that the conceptual framework of commonsense psychology can be made manifest without any empirical or scientific investigation – it is implicit in our everyday practice. This conceptual framework yields a theoretical structure (our commonsense psychological theory of people and how they behave) that specifies functional roles and hence allows us to work downwards from the personal level to the subpersonal level.

Psychological functionalism (or: *psychofunctionalism*) is not as confident as philosophical functionalism about our ability to identify and formulate the theoretical structure that will occupy the top level of the hierarchy of explanation. It asks (quite reasonably) why we should have such confidence in our own understanding of commonsense psychology. Everyday psychological explanations rarely (if ever) explicitly involve subsumption under the sorts of causal generalizations that are supposed to be the stock-in-trade of commonsense psychology. On those occasions when we actually do formulate explicit explanations/predictions of the behavior of others, we typically do no more than cite candidate propositional attitudes without bringing in any generalizations. We might implicitly be presupposing causal generalizations linking those propositional attitudes to the behavior we are trying to explain or predict, but it may well be no easy matter to work out what those causal generalizations are. It may be the case, for example (to present a crude version of a view that is becoming increasingly popular and that we shall examine in more detail in section 8.4), that much of our interpersonal interaction is governed by a social cognition module, the product of a much earlier period of human evolution and consequently sensitive to patterns in behavior that do not correspond to our reflective self-understanding. If this is the case we will not be able to derive a taxonomy of the generalizations of reflective commonsense psychology directly from our everyday explanatory practices – nor, *a fortiori*, by listing all the commonly accepted platitudes we can think of. A process of genuine investigation will be required. And, for the psychological functionalist, the natural conclusion to draw is that this genuine investigation will be the province of scientific psychology.

Psychological functionalism is significantly at odds with philosophical functionalism on its understanding of the aims, scope and explanatory pretensions of scientific psychology. Whereas philosophical functionalism assumes that scientific psychology will uncover causal laws isomorphic to the personal-level causal laws implicated in commonsense psychological explanation, psychological functionalism is skeptical about the role of laws in psychology. At the hands of some prominent psychological functionalists, such as Robert Cummins (Cummins 2000), this skepticism is part of a wholesale skepticism about the explanatory power of what is known in the philosophy of science as the *deductive nomological* (DN) model of explanation (Hempel and Oppenheim 1948). According to the DN model, the paradigm of explanation is the subsumption of an event under a causal law, so that one explains why the event occurs by citing the law-like generalization under which it falls. Cummins argues that the DN model of explanation is fundamentally misconceived, because law-like generalizations simply redescribe the phenomenon that one is trying to explain. Laws, for Cummins, are *explananda* (things to be explained) rather than *explanantia* (things that do the explaining). Laws are useful for prediction, he thinks, but they do not do any explanatory work. Explanation and prediction need to be separated out. Just as we can explain certain things without being able

to predict them (Cummins's example is the swirling trajectory of a falling leaf), so too can we predict things without being in any position to explain them (Cummins here cites the understanding of tide tables that long predated Newton's explanation of the tides).

There is no need to evaluate Cummins's general skepticism about DN explanation. For present purposes we need only note two points that support the psychological functionalist's position. First, as several authors have noted (e.g. Patterson 1996), there are remarkably few laws in psychology. Psychology is just not a good place to look for the sort of causal laws governing the relation between mental events and behavior upon which philosophical functionalism relies. Second, the laws that do exist in psychology can with some plausibility be viewed as *explananda* rather than *explanantia*. We do find certain laws in *psychophysics* (the experimental study of how sensory systems detect stimuli in the environment).[9] But these laws are statistical rather than explanatory. They are confirmed by their instances, rather than explaining their instances. Consider, for example, the Stevens Law in psychophysics, which holds that

$$\Psi = k\Phi^n$$

In this equation Ψ is the perceived intensity of a stimulus and Φ is a physical measure of intensity (e.g. temperature according to some scale), while k and n are constants, with n depending on the type of stimulus (e.g. temperature = 1.6 and electric shock = 3.5). The Stevens Law produces robust predictions of how subjects report the perceived intensity of a range of stimuli. It is hard to see, however, that we are given any *explanation* by being told that the extent to which someone yelps with pain on being burnt is fully in line with what we would expect from the Stevens Law. Instead, according to Cummins and other psychological functionalists, the Stevens Law tracks a robust phenomenon (what they call an *effect*) that requires a fundamentally different type of explanation. We will explore this different type of explanation in the next section.

3.5 Psychological functionalism and the interface problem

Psychological functionalism takes a more cautious view than philosophical functionalism (or the autonomy theory) of how much we know about commonsense psychology. This leads to an even more fundamental difference when it comes to responding to the interface problem. Both philosophical and psychological functionalists think that the key to tackling the interface problem is the notion of realization – the distinction between role and realizer and the concomitant idea that functional roles are multiply realizable.

9 See the annotated bibliography for Chapter 2 for relevant references.

Psychological functionalists object, however, to how philosophical functionalists apply the notion of realization.

Most philosophical functionalists operate with the simplifying assumption that there will be a uniform account of how commonsense psychology is realized in the human nervous system – that there will be a single realizer for each functional role identified through the causal generalizations of commonsense psychology (although the thesis of multiple realizability leaves open the possibility that each functional role might have completely different realizers in different species, and indeed within a given species). Hence we will solve the interface problem in the human case by identifying the relevant realizers. From the viewpoint of psychological functionalism, however, this way of responding to the interface problem is simply too crude. It assumes that there are only two basic levels of explanation – the functional level of commonsense psychology and the subpersonal realization (or *implementational*) level. Yet there are, as we have seen, many different levels of explanation and many different explanatory disciplines at the subpersonal level. Philosophical functionalism seems to collapse them all into one, without any sensitivity to the variety and richness of analysis available at the subpersonal level. The point is put very clearly by William Lycan in the following passage:

> My objection is that "software"/"hardware" talk [or "function"/"structure" talk] encourages the idea of a bipartite Nature, divided into two levels, roughly the physicochemical and the supervenient "functional" or higher-organizational – as against reality, which is a multiple *hierarchy* of levels of nature, each level marked by a nexus of nomic generalizations and supervenient on all those levels below it on the continuum. See Nature as hierarchically organized in this way, and the "function"/"structure" distinction *goes relative*; something is a role as opposed to an occupant, a functional state as opposed to a realizer, or vice versa, only *modulo* a designated level of nature.
>
> (1987, p. 78)

The philosophical functionalist certainly seems guilty of an oversimplified understanding of what goes on when we move below the level of commonsense psychology – not least because the idea of subpersonal psychology discovering a network of generalizations isomorphic to the network of generalizations which make up commonsense psychology seems a basic misunderstanding of the enterprise in which psychologists and neuroscientists are engaged. Psychology (and cognitive neuroscience even more so) is remarkably lacking in law-like generalizations. Relatedly, psychologists and neuroscientists do not see their job as the explanation or prediction of particular instances of behavior. They are more interested in explaining the particular mechanisms that make cognition and cognitively motivated behavior possible (Patterson 1996; Cummins 2000).

These two concerns come together in the psychological functionalist's response to the interface problem. In response to the first perceived shortcoming of philosophical functionalism, the psychological functionalist quite simply denies that resolving the interface problem is a matter of identifying vertical realization relations between the state-types that features as nodes in the network of causal generalizations making up commonsense psychology, on the one hand, and physically identifiable state-types implementing the relevant causal roles, on the other. The vertical explanatory relations are more subtle. Vertical explanation yields an account of the basic cognitive capacities underpinning the horizontal explanations of commonsense psychology. Commonsense psychological explanations attribute particular mental states in the interests of explanation and prediction. These attributions work on the assumption that the basic cognitive capacities of the subject to whom they are made are functioning properly. The job of subpersonal psychology is to explain how these cognitive capacities work.[10] This type of explanation works, moreover, in a way that does justice to the second set of concerns raised about philosophical functionalism. The principal methodology of subpersonal psychology is *functional analysis* – that is to say, the process of explaining a cognitive capacity by breaking it down into subcapacities that can be separately and tractably treated. Each of these subcapacities can in turn be broken down into further nested sub-capacities. As this process of functional decomposition proceeds we will move further and further down the hierarchy of explanation until we eventually arrive (so it is hoped) at the molecular biology of the neuron. The psychological functionalist maintains (as we saw earlier in the quote from William Lycan) that the distinction between functional role and realizers goes all the way down. All the different subpersonal levels of explanation are linked by the realization relation.

A non-psychological example will help to elucidate this conception of functional explanation. Aircraft use gyroscopic instruments to keep track of the rate at which they are turning, of their pitch attitude (the extent to which the nose is pointing up or down) and of their compass heading. Gyroscopes are basically rotating flywheels mounted in a way that allows them to turn freely in one or more directions. Unlike magnetic compasses, for example, which are affected by acceleration and turning and so cannot give reliable readings during those maneuvers, gyroscopes are rigid in space – that is, they remain stable irrespective of how the aircraft is moving around them. The basic functioning of a gyroscope can be broken into several different functional tasks – the sort of tasks that would confront someone setting out to manufacture a gyroscopic instrument. One basic task is obviously to create a casing containing the flywheel that will allow it to keep spinning in the plane of orientation provided that no forces are applied to it. A second basic task is to provide some mechanism for spinning the flywheel, and a

10 Compare the account of weak reduction in section 3.1.

third basic task would be to hook the spinning flywheel up to an indicator gauge which will display the desired information on the basis of changes in the orientation of the spin axis. Corresponding to these three basic tasks are three functional roles that one would expect to see realized in any properly functioning gyroscopic instrument – the flywheel role, the spinning role and the indicator role.

So far, these functional roles have been specified at a very abstract level. Nothing has been said about the sort of mechanisms that might be employed to realize those abstractly specified functional roles. And, as it happens, there are many different mechanisms that will do the relevant jobs. In some gyroscopic instruments the flywheel is set spinning and kept spinning by a small electrical motor. Other gyroscopes work through vacuum systems that draw high-speed air through a nozzle and blow it into grooves machined into the rim of the flywheel. To put it into the terms introduced earlier, the spinning function is multiply realizable. But things do not stop there, of course. The vacuum system, for example, has its own abstractly specifiable functional role – the role of generating a pressure differential that will produce a current of air powerful enough to operate the flywheel. This role can be realized in many different ways. Some vacuum systems use an externally mounted venturi tube that generates the required pressure differential from the dynamic pressure of the air in the slipstream of the aircraft. Other vacuum systems use a vacuum pump. The functional role of a vacuum pump is obvious enough and that functional role can itself be realized in many different ways. Some aircraft use a wet pump, lubricated with engine oil, while others use a dry pump driven off the accessory case of the engine. Nor, of course, does the process of functional decomposition stop here. We can continue specifying functional roles and identifying realizers until we get to the most basic components of the gyroscope – and indeed still further, since even basic physical concepts such as *molecule* and *atom* are functional concepts.

According to psychological functionalism the solution to the interface problem lies in an analogous strategy of functional decomposition, breaking down the core cognitive capacities identified by scientific psychology into ever-simpler capacities in a process that will eventually "bottom out" in capacities and phenomena that, although still functional, are not mysterious in any psychological or cognitive sense. As the process of functional analysis proceeds the mechanisms identified get more and more "stupid" until we eventually arrive at mechanisms that have no identifiable cognitive dimension. The assumption here being, of course, that we will be on fairly safe ground by the time we arrive at, say, the mechanisms allowing electrical signals to be transmitted between neurons. These mechanisms can be understood using the tools of molecular biology and cognate disciplines in a way that does not differ from how we use those tools to understand mechanisms that have no psychological implications.

But how does the process of functional decomposition actually work? We

can get some clear indications from the example we considered in the previous chapter. Marr's analysis of the early visual system bears some of the key hallmarks of a functional analysis. It involves, for example, a process of abstract task analysis, in which the general function of the early visual system (the task of generating a description of the shape and spatial arrangement of objects in the distal environment from a representation of intensity values in the visual array) is broken down into a series of sub-tasks. One such sub-task is the detection of edges and boundaries. Another sub-task is the computation of surface orientation. Each of these sub-tasks is broken down into further sub-tasks. So, for example, one sub-task involved in the detection of edges is the detection of significant intensity changes in the visual array, while the computation of surface orientation can be broken down into the computation of slant (the angle by which a perceived surface falls away from the vertical plane) and the computation of tilt (the direction of slant). Similarly, the functional analysis is not purely a matter of abstract task analysis. The details of the proposed functional decomposition are also determined by evidence from neuropsychology and from psychophysics.

Nonetheless, Marr's analysis of the early visual system is not really a paradigm case of functional decomposition. As we saw when comparing psychological and philosophical functionalism, one of the key ideas of psychological functionalism is that the levels of explanation form a continuum. There are many more levels of explanation than allowed for by Marr's distinction between the algorithmic level and the implementational level. There is no such thing as *the* implementational level. There are many different levels of structured organization in the nervous system, from the level of neural systems to the level of synapses via the level of neural networks. Each of these has some claim to be *an* implementational level. The same will hold, according to psychological functionalists, at the functional level of explanation. Moreover (and this is perhaps the most salient difference between a Marr-style analysis and a canonical functional analysis) functional analysis is not confined to modular processes in the way that Marr's analysis is confined, as I suggested in section 2.1. Marr is committed to the existence of determinate algorithms for the computation of the specific tasks discovered at the functional level of analysis and it seems plausible that such algorithms will only be available for highly specialized and domain-specific types of cognitive processing.

It will be useful to have an example of how functional analysis might be applied in the sphere of higher and non-modular cognitive abilities. Let us consider the bundle of capacities and abilities that are lumped together (in the conceptual framework of commonsense psychology) under the label 'memory'. The phenomenon of memory is a good illustration of functional decomposition, both because it is clearly not a modular process in the strict Fodorean sense (although it may well have modular components) and because it illustrates the range of inputs that are available for a functional analysis. Starting at the top level it seems sensible to break memory down

into three distinct (although of course interrelated) processes. Memory involves *registering* information, *storing* that information and then *retrieving* the information from storage. This three-way distinction is, of course, just the beginning. The interesting questions arise when we start to enquire how those three functions might themselves be performed. For the sake of simplicity I shall concentrate on the function of information storage.

The most basic functional decomposition in theorizing about how information is stored comes with the distinction between *short-term* and *long-term* memory (usually abbreviated STM and LTM respectively). The evidence for this distinction comes from two different sources. One important set of evidence derives from the study of brain-damaged patients. Experimental tests on patients during the 1960s uncovered a *double dissociation* between what appeared to be two separate types of information storage.[11] One patient, known by his initials as K. F., was severely impaired on memory tests that involve repeating strings of digits or words, but was capable of performing more or less normally on tasks that involve recalling material that he had read, recognizing faces, or learning over time to find his way around a new environment (Shallice and Warrington 1980). A diametrically opposed pattern of breakdown (the classical pattern of *amnesia*) was observed in other patients. Patient H. M., for example, was perfectly normal when it came to repeating strings of words or telephone numbers, but profoundly impaired at taking in and using new information (Milner 1966). It looks very much as if there are two different types of information storage involved here, one involving storing information for a relatively short period of time and the other operating over much longer time periods.

The functional decomposition of information storage into short-term and long-term memories is also a natural interpretation of phenomena identified in laboratory experiments on normal subjects. Many memory tasks involve asking subjects to repeat as many words as they can remember in any order from a list of twenty or so unrelated words. Subjects performing these so-called *free recall* tasks typically show two effects. The *recency effect* is that they tend to recall items from the end of the list first, and to get more of these correct – provided that the process of recall starts within a few seconds of the end of the presentation. With delays of more than a few seconds the recency effect disappears. The *primacy effect*, on the other hand, is that slowing down the rate of presentation improves recall of items earlier in the list compared to items later in the list. If the presentation rate is speeded up, then the primacy effect disappears and the recency effect is enhanced. If, on the other hand, a distractor is employed in the interval between presentation and recall (e.g. by asking subjects to count backwards) then even a short delay of 15–20 seconds will result in a success rate of only 10 percent.

The existence of these various effects speaks to the existence of two

11 A double dissociation between two cognitive abilities A and B is discovered when it is found that A can exist in the absence of B and B in the absence of A.

storage systems with different learning characteristics. It is widely thought that the recency effect is a function of recall from STM – which is why it disappears when there is a significant delay between presentation and recall. Information is not retained for long in STM, and retention in STM is a function of rehearsal (i.e. repeating the digits to oneself), which is why performance falls off so drastically during the tasks where the presence of a distractor inhibits rehearsal. Rehearsal is not required for LTM (which accounts for the 10 percent of information retained during the distractor task), but on the other hand the registering of information in LTM is sensitive to the rate of presentation. This is why slowing down the rate of presentation improves recall of items at the beginning of the list – since one assumes that the items from the beginning of the list will be stored in LTM rather than STM. There are many similar effects in laboratory memory tasks that mesh very well with the double dissociation we looked at earlier to support the functional decomposition of memory storage into two distinct processes – STM and LTM.

But how should these two functional components themselves be understood? In the case of STM one influential analysis has suggested a further functional decomposition into a complex multicomponent system. According to the *working memory hypothesis* developed by Baddeley and Hitch (1974), STM is composed of a variety of independent sub-systems. They identify a system whose functional role it is to maintain visual–spatial information (what they call the *sketchpad*) and another responsible for holding and manipulating speech-based information (the so-called *phonological loop*). Both of these sub-systems are under the control of an attentional control system (the *central executive*).

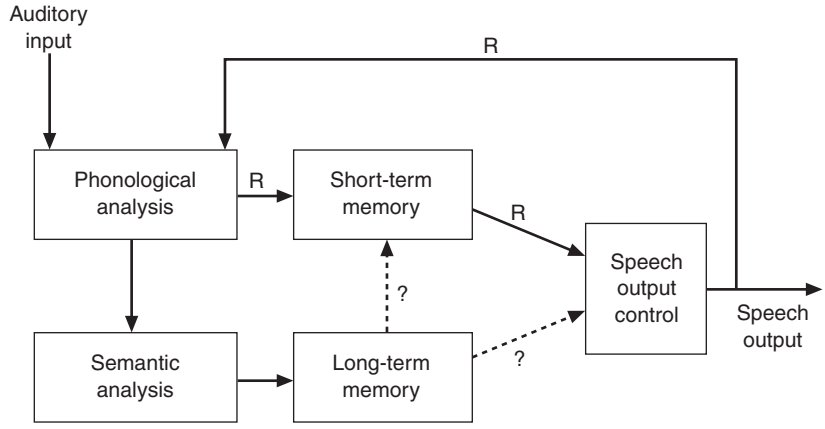

Figure 3.2 Shallice and Warrington's model of the relation between the STM and the LTM involved in auditory-verbal recall. *R* refers to the rehearsal loop (source: adapted from Shallice (1988, p. 55)).

In the case of LTM neuropsychological research has once again been very influential. Evidence from profoundly amnesic patients suffering from *anterograde amnesia* (affecting memory of events after the onset of brain injury, as opposed to *retrograde amnesia*, which extends to events before the injury) has suggested that we need to make a distinction between *implicit* and *explicit* memory systems within the general LTM system. Many such patients have shown normal levels of ability in acquiring motor skills and in developing conditioned responses, even though they have no explicit recollection of the learning process. The tasks on which they perform well are, of course, tasks such as manipulating a computer that do not require the patient to think back to an earlier episode. On tasks of the second type, such as the free recall tasks that we have already briefly looked at, anterograde amnesiacs are profoundly impaired. A further distinction that is suggested by the neuropsychological evidence (and indeed also by experimental evidence from normal subjects) is between episodic memory and semantic memory (Tulving 1972). Episodic memories are directed at temporally dated episodes or events and always have an autobiographical element, while the semantic memory system stores high-level conceptual information, including information about how to speak one's language as well as the various bodies of information that we all possess about the structure of the natural and social worlds.

This is only a very crude sketch of the initial stages of a process of functional decomposition for human memory. But we have enough in front of us to see how the functional analysis might proceed. There are important issues to pursue at both the horizontal and the vertical levels of explanation, even once a preliminary functional decomposition has been made. One question that immediately arises at the horizontal level is: what are the horizontal relations between the various sub-components? This question arises even with the basic distinction between STM and LTM. Should we view STM as a type of antechamber through which information passes on its way to LTM? Or are the systems not just separate but also largely independent of each other? And one might also ask what the inputs are to STM and LTM and what sort of information-processing takes place during the process of registration? Figure 3.2 illustrating Shallice and Warrington's theory of how STM and LTM operate during auditory–verbal recall gives some indication of how these questions might be addressed for a limited type of recall memory.

Functional analysis of this type is sometimes known (often dismissively) as *boxological*. The functional decomposition yields a series of specific functions and capacities, each of which has its own little box. Arrows between the boxes mark the direction of information processing. Critics of "boxology" often suggest that the contents of the boxes (the actual details of how the relevant functions are carried out) are completely mysterious, and hence that a boxological analysis does little more than redescribe the data. It should be clear by now that this is unfair. The process of decomposing functions

(boxes) into sub-functions and further sub-functions is a genuine process of analysis and vertical explanation.

What ultimately makes functional analysis (boxology) genuinely explanatory is that the analysis is discharged at the neural level. And the second set of questions raised even by our brief sketch of a preliminary functional analysis of the memory system is whether the analysis has any implications at the neural level. Does the functional analysis give us any clues as to how it might be anchored at the neural level? As always when we talk about the brain we need to be very cautious, given the limited amounts of information available. But the role that study of brain damage plays in functional analysis offers at least the possibility of anchoring particular cognitive functions in particular brain regions by correlating the locations of brain damage with deficits in particular functions and sub-functions.[12] So, for example, it has been suggested (Baddeley 1998) that the episodic LTM system is located in a circuit linking the temporal lobes, the frontal lobes and the parahippocampal regions. Even at the relatively fine-scaled level of neural implementation (as opposed to larger-scale questions of neural implementation) some suggestions have been tabled. Donald Hebb, a pioneer in research into learning and memory, suggested as long ago as the 1940s that long-term memory storage might involve enduring changes in the patterns of connections among populations of neurons, whereas short-term memory storage might be a matter of patterns of electrical activity in neurons.

Of the approaches to the interface problem so far considered, psychological functionalism is by far the most sensitive to the complex and multi-layered nature of research in scientific psychology, cognitive science and the neurosciences. Whereas the autonomy theorist denies that this research is in any sense directly relevant to commonsense psychology and the understanding of persons, and the philosophical functionalist thinks that it will be possible to find a single realizer for the functional roles identified at the level of commonsense psychology, the psychological functionalist takes seriously the need to integrate the causal mechanisms of commonsense psychology with the practice of research, experiment and investigation at the subpersonal level.

Checklist for the functional mind

- The picture of the functional mind starts off from the idea that explanation at the level of commonsense psychology is causal explanation.

12 Any such process is, however, fraught with danger and difficulties. Not only is brain damage by its very nature a messy and hard-to-quantify process, but there are significant difficulties in identifying and comparing deficits and breakdowns across different patients. Some of these difficulties are discussed in Caramazza (1986) and Shallice (1988, Ch. 1). It is possible to avoid some of these problems by lesioning the appropriate brain areas in monkeys and observing the consequences with tests designed to detect aspects of cognitive functioning analagous to those in humans. But new difficulties emerge with the question of how legitimate it is to make inferences about the functional organization of the human brain from that of the monkey brain.

- Unlike the picture of the autonomous mind, the functional mind understands the causal dimension of commonsense psychology as involving subsumption under causal laws which hold at the personal level.
- Within the overall picture of the functional mind there are two ways of identifying these personal-level causal laws. According to *philosophical functionalism*, these laws can be read off from the platitudes that we are all supposed to accept about mental states and the ways in which they relate to each other and feed into behavior. According to *psychological functionalism*, on the other hand, the laws of commonsense psychological explanation are not so easily accessible. Discovering them will require empirical investigation.
- Philosophical functionalists respond to the interface problem by suggesting that explanation at the subpersonal level will uncover the *realizers* for the mental states that occupy the nodes of the network of causal generalizations that makes up commonsense psychology.
- Psychological functionalists propose to resolve the interface problem by a process of functional decomposition and analysis, identifying the cognitive capacities which underwrite the causal generalizations of commonsense psychology and then breaking those capacities down into sub-capacities and further sub-capacities until we arrive at levels of explanation that are non-cognitive.

4 Causes in the mind

From the functional mind to the representational mind

- Causation by content: problems with the functional mind
- The representational mind and the language of thought
- The mind as computer

In this chapter we turn to the third of the four pictures of the mind that we will be looking at in the first part of this book. This is the representational picture, which construes the mind on the model of a digital computer. The representational approach to the mind has been enormously influential in philosophy, psychology and artificial intelligence. It has been developed in a number of different ways with a number of different motivations. My focus in this chapter will be on the version of representationalism developed and defended by Jerry Fodor, both because it is developed with philosophical problems and issues clearly in view and because it is explicitly focused on the interface problem.[1]

I will be presenting the representational view as a way of addressing some serious issues that arise for the picture of the functional mind, particularly with regard to its claim to do justice to the causal dimension of the mental. These problems are brought out in section 4.1. In section 4.2 I show how the representational picture might be thought to resolve the problems of the functional approach. The final section, 4.3, makes explicit the analogy between the mind and a digital computer that is an integral part of the representational approach.

4.1 Causation by content: problems with the functional mind

The previous chapter explored the dialectic that leads to the functional picture of the mind. A crucial element in that dialectic is the thesis that commonsense psychological explanation is a form of causal explanation. Functionalism (in both its philosophical and psychological forms) can be seen as a proposal for doing justice to the (perceived) causal dimension of commonsense psychological explanation. As we saw, this picture of the mind is a natural consequence of the following three theses:

1 See the annotated bibliography for this chapter for further reading on different approaches.

1 Commonsense psychological explanations are causal explanations.
2 Causal explanations require the existence of causal laws.
3 The causal laws governing commonsense psychological explanation have to hold at the personal level.

We saw that proponents of the autonomous mind generally accept (1), but will reject either (2) or (3). Davidson's version of the autonomous mind accepts that causal explanations require the existence of causal laws, but denies that the causal laws in question hold at the personal level. Supporters of anomalous monism accept (1) and (2) but deny (3). The causal dimension of commonsense psychological explanation is underwritten by generalizations holding over psychological states, but these generalizations hold only when these states are characterized in the language of physical theory, since there are no strict causal laws holding at the personal level. The second strand of the autonomy theory, associated with Jennifer Hornsby and John McDowell, understands the causal dimension of commonsense psychological explanations in a more minimalist way. A particular combination of mental states causally explains a given behavior if and only if it is true that in the absence of that combination of mental states the behavior in question would not have occurred. This is the so-called counterfactual theory of causation, and involves denying (2). Counterfactuals about behavior do not need to be supported by laws.

As we saw in the previous chapter, there is a certain plausibility in the functionalist claim that, *if* commonsense psychological explanation is a form of causal explanation, then causal laws holding at the personal level must underwrite the causal explanations in question. It looks very much as if the counterfactuals implied by ordinary psychological explanations are not brute facts but rather themselves things that require explanation – and the most obvious way of explaining them would be to say that they are a consequence of causal laws. It might well seem unsatisfying to suggest, in the manner of anomalous monism, that the relevant causal laws are not psychological causal laws. If a causal law is to support a causal explanation then it must be formulated in terms commensurable with the terms of the explanation – which is precisely what anomalous monism denies.

But since the appeal of the functional picture of the mind lies primarily in its claim to do justice to the causal dimension of commonsense psychological explanation, it is reasonable to ask whether it provides a fully satisfying account of this causal dimension. Even if one thinks that conditions (1) and (2) imposed by the functional picture are *necessary* conditions for any account that will have commonsense psychological explanations coming out as causal explanations in the desired way, it is still an open question whether they are *sufficient*. The central claim of the representational picture of the representational mind is that the causal dimension of commonsense psychological explanation requires more than simply the existence of law-like generalizations holding over personal-level psychological states. According to

the picture of the representational mind, we need to understand the mechanics of the causal processes that are tracked by commonsense psychological explanations. We need an account of how the mind works that will explain *how* a particular combination of beliefs and desires brings about a particular action. It is not enough, representationalists maintain, to identify a combination of beliefs and desires together with a covering law explaining why that combination of beliefs and desires should, *ceteris paribus*, lead to the behavior in question. This does nothing to address the real puzzle posed by commonsense psychological explanations, which is that they offer explanations and predictions of behavior framed in terms of how agents represent their environment. How, one might ask, can a mere representation (whether a belief, a desire, a hope or a fear) have causal effects within the world?

An ambiguity in the term 'representation' can easily obscure why the causal power of representations might be thought so mysterious. On the one hand, a representation is simply an object like any other – it might be a pattern of sound waves, a population of neurons, a piece of paper or a canvas. Considered in this sense, there is no particular mystery about how a representation can be causally efficacious. We have no difficulty in understanding, for example, how a population of neurons can enter into causal transactions with other populations of neurons. We can see, for example, how a particular pattern of activation in the first population could lead to a further pattern of activation in the second population. And we can even see in principle how a succession of such causal transactions could issue in a particular pattern of bodily movements. But this misses an important element in the notion of a representation. Representations are not just objects like any other. They are things that bear a special *semantic* relation to the world. They possess a *content* that stands for objects or states of affairs extrinsic to them. And it is here that the puzzle lies. The puzzle is not just how representations can have causal effects within the world – but rather how representations can have causal effects within the world as a function of their semantic properties, as a function of the relations they bear to other objects (objects that may not in fact even be in existence). I will be approaching the representational conception of the mind through this puzzle, the puzzle of causation by content.

There is a terminological issue here that needs to be addressed, one that masks a substantive philosophical issue. In this book I am discussing the representational mind as if it were fundamentally different from the functional conception of the mind. There is a sense, however, in which the representational mind is a particular variety of philosophical functionalism, rather than an alternative to it (and this is often how it is presented – see Rey (1997, Chapter 8), for example). The basic distinction between functional roles and the realizers of those roles is just as deeply involved in the representational picture as it is in the two varieties of functionalism considered in the previous chapter. And they both share a deep commitment to the nomological

conception of psychological causality – that is, to the view that mental states explain behavior in ways that are law-governed. From the perspective of the autonomy theory considered in the previous chapter, and indeed from that of the neurocomputational conception of the mind to be considered in more detail in the next chapter, there are far more affinities between the functional mind and the representational mind than there is clear blue water between them. I shall, however, be stressing the differences between the functional mind and the representational mind. One reason for this is that it seems to me that a powerful motivation for the picture of the representational mind is the thought that functionalism does not go far enough in its stated brief of accounting for the causal dimension of the mental. This line of argument is prominent in Fodor (see the taxonomy he proposes in Fodor 1987), but much less prominent in other versions of representationalism. I am proposing it as a way of motivating computationalism in the context of the interface problem, not as an account of why computationalism was originally proposed.

It is characteristic of functionalist approaches to the mind, particularly those that I have termed varieties of philosophical functionalism, to think that the semantic properties of a mental state are determined by the functional role of that state (although there is, as we shall see shortly, a significant ambiguity in this basic idea). That is, the way that a mental state represents the world is fixed by the mental states and non-mental phenomena that give rise to it; by the causal consequences it has for behavior; and by the further mental states to which it gives rise (Loar 1981; Block 1986). Part of the attraction of the distinction between role and realizer is that it allows us to think about the actual physical structure realizing the mental state in question in very indirect terms. We can think about it simply as whatever it is that satisfies the role in question. A functional state is like a black box (to use a popular metaphor). We do not know what is inside the box (what it is that actually realizes the state in question), and nor do we care – provided that, whatever it is, it does what it is supposed to do. There is no need to think about the realizer itself in semantic terms. The realizer enters into certain causal transactions that collectively satisfy the functional role of the state in question, thereby fixing its semantic properties. One implication of this is that the semantic properties of mental states effectively drop out of the picture when we move below the personal level of description, when we consider the realizers of functional roles rather than the functional roles themselves. At the subpersonal level, functional states are realized by physical structures (albeit physical structures understood at a certain level of abstraction) and the causal relations into which those physical structures enter are not a function of their semantic properties. The semantic properties that are so important at the personal level no longer have a role to play at the subpersonal level.

As far as classical versions of functionalism are concerned, this restriction of semantic properties to the personal level is a positive advantage, not least

because it offers the prospect of a reduction of semantic properties to non-semantic properties (Loar 1981; Block 1986). Semantic properties have traditionally been thought to pose significant problems for the project of giving a naturalistic account of the world – that is to say, for the project of showing that all our ways of thinking about the world are in some sense continuous with our scientific ways of thinking about the world, and in particular with the ways of thinking about the world current in the physical sciences. Philosophers concerned with naturalism have tended to see the existence of semantic properties as one of the two main obstacles to a naturalistic picture of the world (the other being the qualitative dimension of certain cognitive states). Correlatively, naturalist philosophers have tended to promote the naturalist agenda by offering reductions of semantic properties to non-semantic properties. Philosophical functionalism offers perhaps the most straightforward way of doing this.

It is natural to ask, however, whether it is possible to give an adequate account of causation through content if one thinks about semantic properties in the manner proposed by the functional picture. And a good way of appreciating the appeal of the picture of the representational mind is via the thought that a proper account of causation through content requires thinking about the subpersonal underpinnings of content in way ruled out by the type of semantics available on the functionalist picture.[2] We can start motivating the picture of the representational mind by thinking about why one might feel that a proper account of causation through content is unlikely to be forthcoming on the functional picture.

We need to be more explicit about what we are looking for in an account of causation by content. There are three principal *desiderata*. The first *desideratum* is an account of how mental states have content. What makes it the case that a particular physical structure, say a particular population of neurons, represents the world in a particular way? The second *desideratum*, following on straightforwardly from the first, is an account of how those representational aspects of mental states can be causally effective. How can the causal properties of a mental state be a function of its content? The third *desideratum* is something we have not yet touched upon. We need, not just a model of how mental states can have contents and be causally efficacious in virtue of those contents, but also a model of how those mental states can feature in a process of thinking. Something needs to be said about how causal interactions between mental states can yield rational transitions between mental states, not to mention rational transitions between mental states and behavior.

The philosophical functionalist thinks that all three desiderata can be

2 This is not to say, however, that the picture of the representational mind is in any sense committed to the irreducibility of semantic properties. Representationalists have their own theories of how to secure a reduction of semantic properties to non-semantic properties. For an overview, see Chapter 9 of Rey (1997). Further reading will be found in the annotated bibliography.

satisfied simultaneously by appeal to the notion of functional role. A mental state has the content it does in virtue of the functional role it plays both within the mental economy and in mediating behavior. This simple idea is at the heart of the different varieties of philosophical functionalism. The core functionalist claim is that the notion of content is to be elucidated through the notion of functional role. Here is how the point is put by a contemporary functionalist:

> The basic functionalist thesis is that psychological state types are to be characterized in terms of the functional roles that those states play in a structure of internal states, mediating stimulus inputs and behavioral outputs. Considerations of functional role are held to differentiate not only among general psychological state types such as beliefs, desires, and intentions, but also among sub-types such as *believing that p* and *believing that q*. A state has whatever content it does on the basis of its functional role.
>
> (Van Gulick 1980, p. 108)

The last two sentences are important. The idea that semantics is fixed by functional role is supposed to apply, not just to types of attitude (that is to say, to beliefs as opposed to desires), but also to particular propositional attitudes (to the particular belief, for example, that Paris is the capital of France). It is standard to analyze a propositional attitude, such the fear that the roof will fall in, into two components – namely, the attitude of fear and the particular proposition to which that attitude is being taken (the proposition that the roof will fall in). A thinker can take different attitudes towards the same proposition at different times (I can believe that the roof will fall in and, after it has fallen in, I can regret that it did so) or, for that matter, at the same time (I can believe that the roof will fall in but secretly hope that it won't). This proposition is what is normally thought of as the content of the relevant attitude. So, the functionalist proposal is that considerations of functional role will be sufficient to identify psychological states even at this fine-grained level of description. The belief that the roof will fall in will have a distinctive functional role that marks it out not simply from the desire that the roof fall in, but also from the belief that the roof is red and indeed from any other belief.

The suggestion that psychological state types are individuated at the level of content by their functional role holds obvious promise for solving the puzzle of causation by content. Whenever a mental state is involved in a particular causal transaction it is, as a matter of definition, exercising its functional role – given that the causal transactions into which it enters effectively define its functional role. The relation between what a mental state does (its functional role) and how it represents the world (its content) is so close that there is no difficulty in seeing how a functionally defined mental state can be causally efficacious in virtue of its content. Content is

determined by functional role. There is no possibility of the two coming apart.

Everything depends upon the plausibility of the thesis that functional role fixes content, and it is not difficult to see why a theorist might be skeptical about this. One important set of worries is best left aside until we consider the extent to which the picture of the functional mind can accommodate the idea of thinking as a process, but that still leaves us with two sources of concern, one practical and the other theoretical. The practical concern is obvious. No one has ever come close to providing an account of the functional role of any content-bearing state. It is no accident that functionalism is most often presented in the context of non-content-bearing states such as pain. It is not hard to see how pain might have a fairly determinate and easily identifiable functional role fixed by its typical causes and typical behavioral effects. But of course what makes pain so straightforward is that it does not, in any obvious sense, involve representations of the world.[3] The problem comes when we try to extend the account to mental states that do involve representations. It is instructive to think back to the example of philosophical behaviorism.[4] According to philosophical behaviorism, psychological states should be viewed as complicated dispositions to behave in certain ways. No philosophical behaviorist, however, ever produced so much as a single comprehensive analysis of a psychological state in terms of such dispositions.[5]

Admittedly, philosophical functionalists (and in particular folk functionalists) are a step ahead of philosophical behaviorists in that they have, as we saw in the previous chapter (see section 3.4), proposed a method for obtaining a functional analysis of mental states. Here again is an important passage from David Lewis, quoted in Chapter 3:

> Think of commonsense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding

3 A conspicuous example here is what is often thought of as the original functionalist "manifesto" – Hilary Putnam's 'The nature of mental states' (Putnam 1967). An increasing number of philosophers have challenged the view that pain is not a representational state, suggesting that pains (and other bodily sensations) do in fact represent events taking place in the body. See Armstrong (1962) for a pioneering effort in this direction and Harman (1990) and Tye (1990) for more recent approaches. It is revealing, however, that none of these philosophers have proposed an account of the semantics of pain (of how pain states actually represent the body) in terms of functional role.

4 Not least because functionalism is often analyzed (with some plausibility) as a causal version of philosophical behaviorism. One might say, oversimplifying somewhat, that whereas the philosophical behaviorist identifies mental states with dispositions to behavior, the functionalist identifies mental states with whatever it is that causes those dispositions to behavior (and, additionally, is caused in certain ways etc.).

5 Even Ryle (1949), the most sophisticated and worked-out formulation of philosophical behaviorism, contains no such comprehensive analysis (although it contains numerous insights into different types of mental state).

> the causal relations of mental states, sensory stimuli and motor responses ... Add also all the platitudes to the effect that one mental state falls under another – "toothache is a type of pain" and the like ... Include only platitudes which are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that the names of mental states derive their meaning from these platitudes.
>
> (1972, p. 212)

But there is room for skepticism about the procedure Lewis proposed. The only platitude Lewis actually gives is a platitude involving non-content-bearing psychological states and it is difficult to see how it can be extended to individual psychological states with particular contents. It is true that there are certain commonly accepted platitudes defined over content-bearing states. So, for example, functionalist authors often (and quite plausibly) maintain that something like the following platitude is a mainstay of commonsense psychology: "If someone desires that *p* and believes that ø-ing is the best way to bring it about that *p*, then, all other things being equal, they will ø." This is certainly a platitude defined over content-bearing states. But it is not a good example of the sort of principle that the functionalist requires, since it is really a schema in the logician's sense. It identifies the common structure of the indefinitely many specific principles that are obtained when one substitutes a particular proposition for *p* and a particular description of an action for ø. The general principle itself can give us no help with working out those particular propositions and descriptions. For that we need to appeal to much more specific functional roles – the functional role, for example, of the belief that St Louis is in Missouri, or the desire that St Louis contains fewer restaurants. Are there really such finely-grained functional roles? And even if there are, how could we possibly go about discovering them? A critic of the functionalist project will insist that the burden of proof is on the functionalist to show that functional roles can actually individuate the content of beliefs and other propositional attitudes. One might wonder, for example, how the functionalist can be so sure that there actually are commonly accepted platitudes associated with each of our beliefs and desires. And, even if there are, what grounds are there for thinking that they will be consistent with each other, or that they will uniquely identify a causal/functional role?

Quite apart from these practical concerns there is a more fundamental worry. Recall that the standard model of the propositional attitudes imposes a sharp distinction between attitude and content – between the particular proposition that is the object of one's mental state and the mental attitude one takes towards it. This is a distinction that goes back to the beginnings of philosophical logic and the philosophy of language in Gottlob Frege's *Begriffsschrift*, and it is a distinction that seems completely indispensable to a proper account of thought and the mind. For one thing, as we have already

seen, much of our understanding of ourselves and others rests upon our being able to make sense of the idea of a particular person being able to take different attitudes to the same proposition at different times (and indeed at the same time), as indeed of different people being able to take different attitudes to the same proposition. Without this we would be unable to make sense of either the fundamental continuities holding across a person's life or basic disagreements between people. Equally importantly, a crucial element in thinking involves entertaining propositions without taking any sort of attitude towards them. The most obvious example occurs in conditional thought, which plays such a central role in practical decision-making. I might, for example, believe that if *A* is the case then I ought to ø without actually believing that *A* is the case. Of course, the conditional "If *A*, then I ought to ø" might feature as part of a *modus ponens* inference, combining with the premise that *A* is the case to yield the conclusion that I ought to ø – in which case I would be believing that *A* is the case. But, on the other hand, the conditional might equally feature in a *modus tollens* inference, with the rejection of the proposition that I ought to ø leading to the rejection of the proposition that *A* is the case. In which case I would be entertaining the thought that *A* is the case without ever believing it.

The distinction between attitude and content is, then, of fundamental importance, but it is very hard to see how any version of functional role semantics can accommodate it. The point is straightforward, although it has not received any attention in debates on functionalism. Contents *per se* do not have functional roles. There are no characteristic causes of a particular proposition, or typical behavioral outputs to which a given proposition will give rise. Propositions stand in inferential relations to other propositions, not in causal relations to perceptions or behavior. A proposition can only acquire a distinctive functional role when a particular attitude is taken towards it. The proposition that the roof will fall in, for example, does not have any implications for how someone entertaining it will behave until that person comes to believe it (or to fear it or to hope for it, or whatever). Simply entertaining the proposition is behaviorally neutral.[6] If we are analyzing at the level of functional role we will have to give separate accounts, for arbitrary *p*, of the belief that *p*, the desire that *p*, the hope that *p*, and so forth. Each of these will (if we bracket the concerns raised earlier) have a distinctive functional role. We will most likely not be able to analyze each of these different propositional attitudes as different ways of relating to a single proposition, in the way that the distinction between content and attitude suggests.

6 There may well be some behavioral implications of entertaining a proposition. Entertaining the proposition that the roof will fall in may make me, for example, more likely to come up with "roof" rather than "root" or "rook" if I am asked to think of a four-letter word beginning with "roo–". I take it, however, that *this* type of behavioral consequence is unlikely to define a robust notion of functional role. My thanks to Fiona Macpherson for this example.

But why should this be a problem? Why can we not take attitudes to be primary and derive contents from them? Why can we not, for example, start off by analyzing the distinctive functional role of a set of full-fledged attitudes and then work backwards from that to the content that they all share? This content would be a proposition that might feature as the antecedent of a conditional and that would serve to explain both intrapersonal and interpersonal psychological continuities in the manner discussed earlier. But there is no reason to think that there will be any such core *at the level of functional role*. There is no reason to think that the analysis of functional role will follow the model of the attitude–content distinction, in a way that would allow the "attitudinal" functional role to be peeled away to yield the "propositional" functional role. Consider my belief that it is raining, to take a standard example. This belief has a fairly clear-cut functional role. Its standard causes include, for example, perceiving that it is raining, or being indoors and hearing on the radio that it is raining. The belief also has fairly obvious effects. It will cause me to believe that the sun will only be shining if there is a rainbow, for example, and it will lead me to put some kind of rainwear on if I decide to go outside and want not to get wet. Consider, on the other hand, my desire that it rain. This also has a fairly clear-cut functional role, but there seems to be no overlap between that functional role and the functional role of the belief that it is raining. My desire that it rain is not in any sense caused by the sort of thing that causes me to believe that it is raining – and nor does it have similar effects. When they are considered purely in terms of their respective functional roles, my belief that it is raining and my desire that it rain have little, if anything, in common – even though they are different attitudes to the same proposition. It seems very likely that this belief–desire pair is entirely typical in this respect. Generally speaking, whereas the belief that p is typically caused by its being the case that p, the desire that p is typically caused by its *not* being the case that p. So, if the idea that content is to be determined by functional role is taken seriously, it looks as if it will turn out that there is nothing interesting in common between different propositional attitudes being taken to a single propositional content. And this, one might think, would be a serious misrepresentation of the nature of thought.

There are two reasons, then, why a theorist might be skeptical about whether the functional picture of the mind can really give a satisfactory account of the semantic properties of mental states in terms of functional role semantics. In addition to practical concerns about the feasibility of identifying functional roles, it is unclear whether such an account could accommodate the key distinction between content and attitude. What about the third *desideratum* identified earlier? Can the functional picture of the mind explain how causal interactions between mental states yield rational transitions between mental states – can it explain how the transitions between mental states are both causal and inferential? The discussion so far has provided at least a *prima facie* reason for thinking that the functional approach

may have some difficulty here. The issue is (once again) the distinction between attitude and content. It is propositional attitudes that feature in causal interactions, yet it is propositions pure and simple that bear inferential relations to each other.[7] If it is right to claim, then, that functional role semantics cannot operate at the level of propositions, then it looks as if the functional picture of the mind will have difficulties here. This will be pursued further in the next section, when we will see how the picture of the representational mind has a distinctive proposal for overcoming these difficulties.

4.2 The representational mind and the language of thought

It emerged in the previous section that there are certain basic requirements upon any explanation of how propositional attitudes can be causally responsible for generating behavior (or, for that matter, other propositional attitudes). It is not enough to explain how mental states can be causally efficacious. We need to know how those mental states can be causally efficacious in virtue of their content – in virtue of how they represent the world. Moreover, the causal dimension of mental states is not simply a matter of how mental states generate behavior, but also a matter of how they can combine inferentially to generate further mental states.

The picture of the functional mind proposes to tackle the problem of causation by content by analyzing the content of a propositional attitude in terms of the functional role of the physical structure that realizes it – that is to say, in terms of the causal transactions into which that physical structure typically enters. The problem of causation by content is solved because whenever a mental state is involved in a particular causal transaction it is, as a matter of definition, exercising its functional role – given that the causal transactions into which it enters effectively define its functional role. The functional picture of the mind does not allow a gap between what a mental state does (its functional role, as given by the functional role of its realizer) and how it represents the world (its content). But, as we saw in the previous section, there are potential difficulties with the idea that the content of a mental state can be given in terms of its functional role. Not only are significant and quite possibly insuperable practical difficulties in giving plausible functional accounts of particular propositional attitudes, but the functional approach has difficulties accommodating the all-important distinction between content and attitude (and hence in explaining how different propositional attitudes can have the same content).

7 This point has been emphasized by Gilbert Harman in various places (curiously, since he is an influential proponent of functional role semantics). As Harman has stressed, there is a fundamental distinction between logic (which has to do with the relations holding between propositions) and reasoning (which is a dynamic process that involves the evolution and interaction of propositional attitudes). See, for example, Harman (1973, 1999).

The representational picture as I am presenting it is both a development of the functional picture and a departure from it. The representational picture does employ the notion of functional role but is not committed to employing it as globally as the functional picture. The representational picture shares with the functional picture the view that all propositional attitudes are realized by physical structures that have a particular functional role in virtue of the causal transactions into which they enter. But the representational approach does not have to hold that the semantic properties of an individual propositional attitude are exhausted by the functional role of the physical state that realizes that attitude. Representational theorists can take functional role to determine semantics only to the extent of determining which particular attitude is in play.[8] Consider, for example, a particular physical structure that realizes a particular propositional attitude – say, the belief that *p*. According to the functional picture, what makes this physical structure the realizer of the belief that *p* is that it enters into the causal interactions constitutive of the belief that *p*. It is open to the representational theorist, on the other hand, to hold that the causal interactions into which that physical structure enters only determine that it is a belief, as opposed, for example, to a desire. Functional role determines the particular attitude, but not the particular content.

There is a familiar metaphor often used by proponents of the picture of the representational mind. They talk about a particular token mental state being in the "belief box" or the "desire box". Talk of belief boxes and desires boxes is intended to convey the idea that beliefs and desires have separate functional roles. And talk of mental states being "in" one box rather than another reinforces the idea that mental states have the content that they have independently of their functional role. The representational picture thereby avoids the two problems that threaten the functional picture. It has built into it precisely the type of distinction between content and attitude that the functional picture finds difficult to accommodate and it is not in any sense committed to the implausible idea that the content of a propositional attitude can be determined by its functional role.[9]

What makes it possible for representational theorists to side-step these difficulties? This brings us to what is really distinctive in the picture of the representational mind, which is how it understands the relation between a

8 There are representational theorists who combine a version of the language of thought hypothesis with a version of functional role semantics. Gilbert Harman is a prominent example (see his 1987). Harman's version of functional role semantics is distinctive in being a theory of concepts (the constituents of propositions) rather than a theory of propositions. Whereas a standard functionalist semantics takes propositional attitudes to have functional roles, Harman's functional role semantics is based on the inferential roles of individual concepts. To the extent that he operates within a theory that stresses the internal structure of propositional attitudes, his position comes closer to the computational picture than to the functional picture.

9 Of course, if the picture of the representational mind abandons functional role semantics (which it need not do – see n.8 above), then it incurs the obligation to give an alternative account of how

particular propositional attitude and the physical structure in which it is realized. What distinguishes the representational picture, as I understand it, is that it takes the physical realizers of propositional attitudes to be internally structured. No such internal structure is required by the functional picture of the mind. It is individual attitudes that are standardly taken to have functional roles. The notion of functional role typically is not applied below the level of the proposition. As Fodor puts it (1987), construing the semantics of propositional attitudes in terms of functional roles goes hand in hand with taking propositional attitudes to be *monadic*. Let us look in more detail at how propositional attitudes are supposed to be structured on the representational approach.

Propositional attitudes, as they are usually understood, have contents that can be specified by sentences following 'that–' clauses. So, for example, if Isolde believes that Tristan has drunk the death potion then the content of her belief is given by the sentence "Tristan has drunk the death potion". On most construals, the content of the attitude has a structure that is isomorphic to the structure of the sentence expressing it. So, to continue with the example, Isolde's belief that Tristan has drunk the death potion is composed of distinguishable components that correspond to the distinguishable components of the sentence expressing its content – a component corresponding to the proper name 'Tristan', a component corresponding to the definite description 'the death potion' and a component corresponding to the relational predicate '– has drunk–'. These components can feature in different thoughts (taking 'thought' and 'proposition' to be synonymous). The component corresponding to "Tristan", for example, features in the thought that would be expressed through the sentence 'Tristan is behaving strangely'. The component corresponding to the definite description 'the death potion' features in the thought that would be expressed through the sentence "Brangäne has brought the death potion". The fact that thoughts are composed of distinguishable components that can feature in further thoughts plays an important role in explaining how thoughts can be inferentially connected. We can follow the standard usage by calling these distinguishable components concepts.

There is an important sense, then, in which the contents of propositional

---

mental states have content. Whereas the account of mental content given by philosophical functionalists is derived from the functional role of types of mental state, the representational picture goes most naturally with a *relational* account of mental content. Relational accounts of mental content are based upon the relations that hold between particular types of mental state and the objects or properties represented by those types of mental state. The simplest type of relational account derives the content of particular types of mental state from the typical causes of tokens of the relevant type. More sophisticated accounts stress causal covariance rather than causation. A third type of account attempts to derive semantic properties from the function of mental states. The pros and cons of these different approaches will not be discussed in this book. A brief overview of the principal theories in the market will be found in Chapter 9 of Rey (1997) and in Loewer (1997). Guidance on further reading will be found in the annotated bibliography for this chapter.

attitudes are *structured*. But there are two different ways of thinking about the structure of propositional attitudes. Consider a belief – the belief, say, that La Paz is the capital of Bolivia. The content of this belief is the proposition that La Paz is the capital of Bolivia – a proposition that might be expressed by the sentence "La Paz is the capital of Bolivia" or by a translation of that sentence into Spanish or any other language. As we have seen, we can view that proposition as being structured by viewing it as composed of parts that more or less correspond to the parts of the sentence "La Paz is the capital of Bolivia". But, according to both supporters of the functional picture of the mind and proponents of the representational mind, there is more to a belief than its content. The analogy with written and spoken language is instructive. Consider the written sentence "La Paz is the capital of Bolivia". The inscriptions on the page serve as the vehicle for the proposition that this sentence expresses – just as a complex pattern of sound waves serves as the vehicle for the very same proposition when the sentence is uttered. The standard view is that when an individual believes that La Paz is the capital of Bolivia, the content of a belief is realized by a physical structure that is the *vehicle* of its content in just the same way as the meaning of the sentence is realized by the pattern or sound waves or the inscription on the page.

When the distinction between vehicle and content is made explicit, it becomes clear that there are two ways that a belief, or any other propositional attitude, might be structured. It might be structured at the level of content or at the level of vehicle. We have so far been discussing how the *content* of a belief might be structured. But what separates out the picture of the representational mind from the functional picture is the question of whether beliefs with structured contents are realized in physical vehicles that themselves possess an internal structure – that is to say, the question whether there must be a structural isomorphism between the content of a belief and the vehicle of a belief. No such structural isomorphism is envisaged within standard developments of the functional picture of the mind. As we saw in the previous section, a mental state is understood on the functional picture as the occupier of a causal role, and that causal role determines the mental state's content. But the causal role does not itself possess a structure. According to the functional picture, what makes a physical structure the vehicle of my belief that La Paz is the capital of Bolivia is the causal role of that structure – the things that give rise to it and the things that it typically leads me to do. This causal role is a complex phenomenon, in the sense that it is it is made up of a great number of behavioral dispositions and causal tendencies. But it is not structured in anything like the way in which the content of the belief is structured. It does not, for example, contain an identifiable component corresponding to Bolivia, or one corresponding to La Paz. The content of the belief is structured, on the functional picture, but it does not have an isomorphically structured vehicle.

The distinguishing feature of the picture of the representational mind, in

contrast, is the claim that we can only understand the causal dimension of propositional attitudes through taking them to have structured vehicles – more precisely, vehicles with a structure isomorphic to that of the contents of those propositional attitudes. In order to explore this further we need a clearer view of the overall position within which it is embedded. The central idea of the representational view, as we find it developed by Jerry Fodor, its most prominent exponent, is that propositional attitudes are relations to sentences/formulae in an internal language of thought. These sentences/formulae represent the thought in a way that allows its content to be deployed in reasoning and decision-making. They serve as inner surrogates standing in for the thought's propositional content in a way that allows that content to be causally relevant (thus solving the problem of causation by content). It is useful to break the picture of the representational mind down into three basic claims, as follows:

1 The causal dimension of propositional attitudes must be understood in terms of causal interactions between physical structures.
2 These physical states have the structure of sentences and their sentential structure governs both their composition and their combination.
3 The causal transitions between physical states respect the rational relations between the thoughts that those physical states represent – as a function of the *intrinsic* properties of those physical states.

Let us look at each of these three claims in turn.

The first claim is common currency among almost all the ways of looking at the mind that we have been discussing. It is disputed only by some versions of the autonomy picture. It would not be accepted (for obvious reasons) by those extreme autonomy theorists who think that it is some sort of category mistake to describe mental states as entering into causal relations. Nor would it be accepted by those autonomy theorists who think that the causal dimension of mental states must be understood in terms of counterfactuals about how the person in question would have behaved in different circumstances. But these are both minority views. The first claim, therefore, is relatively uncontroversial. The distinctiveness of the representational view emerges with the second claim – with the notion that the physical vehicles of propositional attitudes have the structure of sentences. We have already looked at the basic distinction between structure at the level of content and structure at the level of vehicle. The issue now is the precise type of structure proposed at the level of the vehicle.

The relation between structure at the level of content and structure at the level of vehicle was introduced earlier by analogy with the structure of a natural language sentence and the structure of the proposition it expresses. In fact, this is more than simply an analogy, since the picture of the representational mind holds that the vehicles of propositional attitudes really are sentences. They need not, however, be sentences in a natural or public

language, but rather could be (and on Fodor's view are) sentences in an internal and private language of thought. It is important not to take talk about sentences in an internal language of thought too literally. The suggestion is not that we will find some sort of mysterious language-like inscriptions if we look hard enough at cerebral matter with acute enough instruments. The hypothesis is formulated at Marr's algorithmic level (to return to the distinction between levels of explanation explored in section 2.1), not at the implementational level, and so it is formulated at a level of abstraction from the physical details of what takes place in the brain. The claim is that, at the level of analysis and abstraction at which it is appropriate to think about the causal transitions into which propositional attitudes can enter, we need to view those attitudes as being realized in physical vehicles that have the structure of sentences. But what is it for a physical vehicle to have the structure of a sentence? It is for there to be a structural isomorphism between components of the vehicle and components of the sentence expressing the content of the propositional attitude in question.

The notion of a structural isomorphism here needs to be understood as imposing two different requirements. The first is that it should be possible to identify in the vehicle of, say, my belief that La Paz is the capital of Bolivia, distinguishable physical elements corresponding to the basic elements of the sentence "La Paz is the capital of Bolivia". The basic elements of the sentence are its *semantically* basic elements, the concepts that it involves, as opposed for example to the letters of which it is composed. We can view these distinguishable physical elements as symbols that stand as surrogates for the semantically basic concepts in the proposition in question. And the description of these physical elements as *distinguishable* means that it is possible for them to appear in other physical structures serving as the vehicles of other propositional attitudes – just as the expression "– the capital of –" can appear in a range of other sentences expressing a range of further beliefs. This is made possible by the second requirement built into the notion of structural isomorphism. The physical elements of which the vehicle of a propositional attitude is made up are combined in ways that map onto the ways individual concepts combine to make up a proposition or a thought. The best model we have for understanding how individual physical symbols standing in for concepts can be combined to form complex symbols standing in for complete thoughts (complete propositions) comes of course from our understanding of language. The conclusion drawn by proponents of the picture of the representational mind is that the vehicles of propositional attitudes are sentences in an internal language. In fact, the conclusion is inescapable once it is granted that there must be a structural isomorphism between content and vehicle in the two senses just described.

The language of thought (LOT) does not have to be a natural language, such as English or Swahili. In fact, understanding the LOT in this way forecloses on at least one explanatory role that the language of thought hypothesis has been called upon to play. As we will see in more detail in Chapter 10,

the language of thought hypothesis has been used to explain how linguistic comprehension is possible. There is a single basic idea common to the principal contemporary approaches to the study of natural languages. This is that the understanding of a sentence proceeds via a grasp of its structure at a level that is independent of, although obviously in some way derived from, its surface syntactical structure (Harman 1972). This structure of a sentence is often called its *logical form*. Since the general model of linguistic comprehension espoused by defenders of the language of thought hypothesis is one on which a natural language sentence is understood by being translated into a sentence in the language of thought, it is clear that the language of thought cannot admit a distinction between surface structure and deep structure. There can be no interpreter who will be able to abstract away from surface syntactical features to identify the deep structure of a sentence in the language of thought. Nor is there any further language into which such a sentence can be translated. So it seems that the language of thought will be much closer to a formal language in being free of the imprecision, ambiguity and vagueness characteristic of natural languages.

We can return to the earlier discussion of the distinction between attitude and content to move towards an overall picture of the mechanics of cognition as envisaged by proponents of the representational mind. Cognition, on the representational picture, is essentially a matter of the generation of propositional attitudes from perceptual inputs and of combining propositional attitudes to generate further propositional attitudes and, ultimately, behavior. Each propositional attitude, each belief and each desire, is to be understood in terms of its particular content – in terms of how it represents the world, either as it is taken to be (in the case of a belief) or as it is desired to be (in the case of desire and other motivational attitudes). The content of a given propositional attitude is realized by a physical structure that is a sentence in the language of thought. That is to say, each content has as its vehicle a complex symbol that is structured in a manner isomorphic to how the content it realizes is built up from individual concepts. What makes it the case that a particular attitude is taken to a particular content is the functional role that the vehicle of that content plays within the overall cognitive economy (the "box" that it is in). By the same token, in virtue of its vehicle occupying different functional roles (either at different times or at the same time), a given propositional content can feature in a range of different propositional attitudes. I can hope that the cat is on the mat, fear that the cat is on the mat, desire that the cat be on the mat or believe that the cat is on the mat.

We see, therefore, two fundamental differences between the representational picture and the functional picture. First, the representational picture is able to make precisely the sharp distinction between content and attitude that is not available on the functional picture. Second, the representational picture identifies structure at the level of the vehicles of content in a way that the functional picture does not. There remains, however, an aspect of

the representational view upon which we have not yet touched. This is its account of the mechanics of thinking over time. To understand how thinking takes place, it is not enough to understand how a belief or a desire is realized in the central nervous system at a time. We need also to understand the mechanics of how transitions between propositional attitudes take place, and of how propositional attitudes can generate behavior. The functional picture of the mind has a relatively simple account of what might be termed the *diachronic mechanics of thinking*. Propositional attitudes are realized by physical structures that interact causally with physical structures realizing other propositional attitudes. These causal interactions constitute thinking. What makes it the case that one belief (say, the belief that *p*) is inferred from two others (say, the beliefs that *q* and *r*) is that the physical structures realizing the beliefs that *q* and *r* jointly cause the physical structure realizing the belief that *p*. The picture of the representational mind takes over this basic model but reinterprets it in the light of the structural isomorphism that it identifies between content and vehicle.

Recall the earlier suggestion that an account of causation by content must incorporate an account of how causal interaction between mental states can yield rational transitions between mental states, not to mention rational transitions between mental states and behavior. It should be easy to see why a theorist might think that this requirement is not met by the functionalist account. The functionalist simply assumes that the causal interactions into which a mental state enters as a function of its causal role will track the rational relations between mental states. The notion of the functional role of a belief does not really make sense unless we assume, broadly speaking, that if it is part of the causal role of the belief that *p* that it cause the further beliefs that *q* and *r*, then it must in some sense be rational to believe that *q* and to believe that *r* if one believes that *p*. This might be because *p* entails *q* and *r* – or it might be because it makes them more likely. Either way, the causal role of a belief must reflect (some core of) the rational relations in which it stands to other beliefs and other mental states. This is something that the functional approach assumes, but not something that it can explain. Why should causal interactions between physical structures track rational connections? Why should my belief that it is raining cause me to have further beliefs that would be true (or likely to be true) if it were indeed raining, if these beliefs are understood as they are on the functional picture? It is very unclear that the functional picture provides the resources to answer this pressing question.

The representational picture, in contrast, has a bold proposal to explain how causal relations between physical structures can track the rational relations holding between the propositional contents that they realize. The key idea is that, since the vehicles of propositional attitudes are complex symbols that form sentences in an internal language of thought, we should understand the relation between vehicle and content in the language of thought on the model of the relation between syntax and semantics in a formal

system. Some background will be useful in explaining how this works and how it is supposed to resolve the problem of explaining how causal relations between vehicles of propositional attitudes can track rational relations between the contents of those attitudes.

The most important feature of any formal system is the clear separation it affords between syntax and semantics, and hence the possibility of viewing the system in two different ways. Viewed syntactically, a formal system is a set of symbols of various types together with rules for manipulating those symbols according to their type. So, for example, the predicate calculus can be viewed as a set of symbols whose role in the system is identified by various typographical features (such as upper case for predicate letters and lower case for individual constants) and that can be combined to make complex symbols according to certain rules that identify the symbols only in terms of their typographical features. An example would be the rule that the space after an upper case letter (e.g. the space in '*F*—') can only be filled with a lower case letter (e.g. '*a*'). Simplifying somewhat, this rule is a way of capturing at the syntactic level the intuitive thought that properties apply primarily to things, but it does this without adverting at all to the idea that upper case letters serve as the names of properties while lower case letters serve as the names of things. It is a matter purely of the *syntax* of the language. The connection between the formal system and what it is about, on the other hand, comes at the level of *semantics*. It is when we think about the semantics of a formal language that we assign objects to the individual constants, properties to the predicates and logical operators to the connectives. To provide a semantics for a language is to give an interpretation to the symbols it contains – to turn it from a collection of meaningless symbols into a representational system.

Just as one can view the symbols of a formal system both syntactically and semantically, so too can one view the transitions between those symbols in either of these two ways. The rule of conjunction elimination in the propositional calculus, for example, can be viewed either syntactically or semantically. Viewed syntactically the rule states, effectively, that if on one line of a proof one has a formula of the form '*A* & *B*', then one can write either '*A*' or '*B*' on the next line. Viewed semantically, on the other hand, the rule of conjunction elimination tells us that if a conjunction is true, then so too will both of its conjuncts be true. All transitions in formal systems can be viewed in these two ways, either as rules for manipulating essentially meaningless symbols or as rules determining relations between the truth-values of propositions. It is because of this that it is standard to distinguish between two ways of thinking about the correctness of inferential transitions in formal systems. From a syntactic point of view the key notion is *derivability*, where one symbol is derivable from another just if there is a sequence of legitimate formal steps that lead from the second to the first. From the semantic point of view, however, the key notion is *validity*, where an argument is valid just if there is no way of interpreting its premises and

conclusion such that the premises are all true and the conclusion false. Put very crudely, an argument is a derivation just if every step follows the rules, while it is valid just if it preserves truth (that is, just if it never leads from a true premise to a false conclusion).

What has this got to do with the language of thought? The connection is beautifully simple. We have seen that the picture of the representational mind takes the vehicles of propositional attitudes to be complex symbols in an internal language of thought. The essence of the representational picture is the suggestion that this way of viewing the vehicles of propositional attitudes allows the relation between vehicle and content to be understood on the model of the relation between syntax and semantics in a formal system. Sentences in the language of thought can be viewed purely syntactically, as physical symbol structures composed of basic symbols concatenated according to certain rules of composition. Or they can be viewed semantically in terms of how they represent the world (in which case they are being viewed as the vehicles of propositional attitudes). And so, by extension, transitions between sentences in the language of thought can be viewed either syntactically or semantically – either in terms of formal relations holding between physical symbol structures, or in terms of semantic relations holding between states that represent the world.

Putting the matter in these terms allows us to reformulate the question of how it can be the case that causal transitions between the vehicles of propositional attitudes reliably track the rational relations holding between the contents of those attitudes. Suppose we think that the causal transitions holding between sentences in the language of thought are essentially syntactic, holding purely in virtue of the formal properties of the relevant symbols irrespective of what those symbols might refer to. Then what we are effectively asking is, What makes it the case that the syntactic relations holding between sentences in the language of thought should map onto the semantic relations holding between the propositional contents corresponding to those sentences? And, if we take seriously the idea that the language of thought is a formal system, then this question has a perfectly straightforward answer. We can expect syntactic transitions between sentences in the language of thought to track semantic transitions between the propositional attitudes that they realize for precisely the same reason that we can expect syntax to track semantics in any properly designed formal system. Proponents of the representational approach to the mind can appeal to well-known results in meta-logic (the study of the expressive capacities and formal structure of logical systems) establishing a significant degree of correspondence between syntactic derivability and semantic validity. So, for example, it is known that the first-order predicate calculus is sound and complete. That is to say, in every well-formed proof in the first-order predicate calculus the conclusion really is a logical consequence of the premises (*soundness*) and, conversely, for every argument in which the conclusion follows logically from the premises and both conclusion and premises are formulable in the first-

order predicate calculus there is a well-formed proof (*completeness*). Put in the terms we have been employing, if a series of legitimate and formally definable inferential transitions leads one from formula *A* to a second formula *B*, then one can be sure that *A* cannot be true without *B* being true – and, conversely, if *A* entails *B* in a semantic sense, then one can be sure that there will be a series of formally definable inferential transitions leading from *A* to *B*.

Of the two notions of soundness and completeness, it is clear that the first is indispensable for any formal system. Nothing could be more useless than a formal system in which legitimate inferential transitions will take one from truth to falsity, thus allowing derivable arguments with true premises and false conclusions. And in fact many formal systems are sound but not complete. Gödel's incompleteness theorem shows that this will be the case for any formal system sufficiently strong to represent arithmetic. But, for present purposes, the important point is that is to take it to have at least some analog of the meta-logical property of soundness. And this in turn means that we can expect its syntax to map onto its semantics in at least the sense that every syntactically derivable transition will be semantically valid. This, it is conjectured by proponents of the representational picture, is the key to solving the problem of causation by content.

The overall contours of the proposed solution come across very clearly in the following passage from Jerry Fodor:

> Here, in barest outline, is how the new story is supposed to go: You connect the causal properties of a symbol with its semantic properties *via its syntax*. The syntax of a symbol is one of its higher-order physical properties. To a metaphorical first approximation, we can think of the syntactic structure of a symbol as an abstract feature of its shape. Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could be environments in which the causal role of a symbol correlates with its syntax. It's easy, that is to say, to imagine symbol structures interacting causally in virtue of their syntactic structures. The syntax of a symbol might determine the causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open.
>
> But now, we know from modern logic that certain of the semantic relations among symbols can be, as it were, 'mimicked' by their syntactic relations: that, when seen from a very great distance, is what proof theory is about. So, within certain famous limits, the semantic relation that holds between two symbols when the proposition expressed by the one entails the proposition expressed by the other can be mimicked by syntactic relations in virtue of which one of the symbols is derivable from the other. We can therefore build machines that have, again within famous limits, the following properties:

- The operations of the machine consist entirely of transformations of symbols;
- In the course of performing these operations, the machine is sensitive solely to syntactic properties of the symbols;
- And the operations that the machine performs on the symbols are entirely confined to altering their shapes

> Yet the machine is so devised that it will transform one symbol into another if and only if the propositions expressed by the symbols that are so transformed stand in certain *semantic* relations – e.g. the relation that the premises bear to the conclusion in a valid argument.
>
> (1987, pp. 18–19)

One thing Fodor stresses in this passage is how the representational picture rests upon a particular conception of the mind as a digital computer. This has been in the background of the discussion so far, but has not yet been made explicit. It will be the subject of the next section.

4.3 The mind as computer

As we have already seen, and as Fodor brings out very explicitly in the passage quoted at the end of the previous section, the proposed solution to the problem of causation by content is very closely bound up with a particular conception of cognitive architecture – and in particular with the thesis that the mind should be understood as a representational device carrying out formally specifiable operations. In this section I will explore this conception of cognitive architecture in more detail (see the annotated bibliography for suggestions for further reading).[10]

The conception of the relation between syntax and semantics that is at the heart of the representational view is also at the heart of the modern theory of computing. The single idea at the core both of computer design and of the picture of the representational mind is that a physical mechanism can make purely syntactic calculations and that inferences over items that

10 It is at this point that the philosophically motivated representational theory of mind intersects with the computational or symbolic paradigm in cognitive science and artificial intelligence. The defining claim of computationalism in cognitive science (as presented, for example, in Newell and Simon 1976, and Pylyshyn 1980, 1984) is that cognition should be understood and modeled in terms of formal operations on syntactically structured representations. In this respect there is little, if any, divergence between computationalism and the representational mind as I have presented it and as it is presented by Fodor. There are differences, however, at the level of motivation. Few computationalists are motivated by the concerns about commonsense psychology that are highlighted by Fodor (and indeed some theorists, such as Stich 1983, have combined computationalism with eliminativism about the propositional attitudes). From a taxonomic point of view, the picture of the representational mind should probably be viewed as a species of computationalism. Clearly, working within the framework set by the interface problem, it is the version of computationalism most relevant to our concerns. The annotated bibliography gives guidance on further reading.

can be given a syntactic interpretation without in any sense adverting to that semantic interpretation. Physical mechanisms can be semantically blind and yet succeed in successfully tracking semantic relations. This basic idea is supported by the meta-logical results briefly discussed in the previous section about the relation between the syntax and the semantics of formal systems – and in particular by the discovery that the first-order predicate calculus is sound and complete. But the meta-logical results alone are not sufficient to motivate either the picture of the representational mind or the theory of computing. What is needed in addition is an account of how syntactic transformations might actually be implemented in a physical system, such as a computer or a mind. Even if we accept that syntax can mirror semantics, this still does not explain how we are supposed to get a brain or a machine to do syntax. How are we to get syntactic transformations between symbol structures from causal interactions between physical structures?

The crucial notion in understanding the final piece of the puzzle here is the notion of *effective computability* – and in particular of an effectively computable function. In brief, an effectively computable function is one that can be computed purely mechanically. One way of thinking about what it is to compute a function purely mechanically is in terms of the notion of an *algorithm* (already considered in the context of Marr's taxonomy of levels of explanation in section 2.1). An algorithm is a finite set of rules that are unambiguous and that can be applied systematically to an object or set of objects to transform it or them in definite and circumscribed ways. The instructions for programming a video recorder, for example, are intended to function algorithmically so that they can be followed blindly in a way that will transform the video recorder from being unprogrammed to being programmed; to switch itself on and switch itself off at appropriate times. So too, to take a more complicated example, are the rules of differentiation by hand that can be applied to an equation to yield its derivative. Here we have a series of rules, together with a definite order in which they are to be applied, that will reliably transform one object (the original equation) into another object (the equation of the derivative). We can say, therefore, that an effectively computable function is any function for which an algorithm can be given.

As should be clear, the notion of effective computability is not a technical notion. That is to say, it is not a notion that can be given a precise and determinate meaning that will settle unambiguously for any case whether it applies or not. Nonetheless, one might ask how this informal notion of effective computability maps onto technical notions that can be given a precise and formal characterization. Is there anything that all effectively computable functions might have in common? Could there be any sort of formal analog of the notion of effective computability? The celebrated Church–Turing thesis claims that all effectively computable functions do have something in common, namely, that they can all be carried out by a simple mechanism known as a Turing machine. Turing machines are beautifully simple. A

Turing machine consists simply of an infinitely long piece of tape divided into cells, in each of which one of a range of symbols can be inscribed, together with a machine head that is capable of reading the symbols in any cell of the tape. At any given moment the machine head is located over one of the cells and is capable of carrying out a limited number of operations. It can read the symbol in the cell; delete the symbol in the cell; write a new symbol in the cell and move one cell to the left or right. It can, of course, do any, some or none of these things depending on its instructions. Any individual Turing machine will have a set of instructions (its *machine table*) determining what it will do as a function of the particular symbol inscribed in a particular cell. The Church–Turing thesis is that every effectively computable function is Turing-computable (where a Turing-computable function is one that can be computed by a Turing machine). In fact, it turns out, due to what is known as Turing's theorem, that it is possible to specify Universal Turing Machines that can mimic any individual Turing machine. The Church–Turing thesis then becomes the thesis that a Universal Turing Machine can compute every effectively computable function. Of course, since the notion of effective computability is an informal notion, there is no sense in which the Church–Turing thesis could possibly be proven. But it is almost universally accepted among logicians, computer scientists and mathematicians.

Going back to our original characterization of an effectively computable function, the algorithm required to compute the function can be identified with the machine table of the Turing machine. In what sense is this making any progress over our original characterization of an algorithm as a set of unambiguous instructions? The important point is that the machine table of a Turing machine is completely blind to the semantic properties of the symbols over which it is defined. The machine table gives instructions for how to transform those symbols. Given the appropriate machine table, these transformations will generate transformations that are syntactically valid within a particular formal system. Moreover, and this is the crucial point, the instructions in the machine table can themselves be implemented by a physical system that is not itself a formal system, or any other type of representational system. The implementation of the instructions in the machine table simply requires a physical mechanism that can scan the cells on the tape and respond appropriately to its internal states, by writing on the tape and/or moving one cell to the left or right. This means that the syntactically describable operations that the Turing machine is carrying out are not the bottom level of analysis. The bottom level comes with the simple mechanical operations of the Turing machine.

The operation of a Turing machine gives us a way of understanding how syntactic operations can be effected in causal terms. The causal operations of the mechanism (the causal connection between, for example, registering a '1' in the operative cell and moving one square to the left) are ultimately responsible for the syntactic processing. There is, correspondingly, a level of

analysis of what is going on in the Turing machine that does not in any way advert to the syntactic operations that are being carried out (let alone to the semantic relations being mimicked by those syntactic operations). This is the level of description that characterizes the Turing machine purely in terms of its physical make-up. To return to the three levels of explanation proposed by Marr and discussed earlier in Chapter 2, the Turing machine illustrates how a particular representational algorithm can be implemented. It provides an illustration of how non-representational causal mechanisms can carry out syntactic operations – and indeed, an illustration of how non-representational causal mechanisms can carry out every syntactic operation that is effectively computable.

So what then is the mind on the representational approach? In one very clear sense the mind is being understood as a computer – as a physical mechanism whose causal operations effect syntactic operations on complex symbols. These complex symbols are the vehicles of propositional attitudes and the syntactic operations defined over them mimic the semantic relations holding between the contents of those attitudes. But it is important not to exaggerate the point. The representational approach does not have to be put forward as an account of every aspect of the mind (although, as a matter of fact, it sometimes is, particularly by some working within artificial intelligence and computer science). The prime motivation for the representational picture, as we have so far considered it and as it is presented by such theorists as Fodor, is with solving a form of what in Chapter 3 I termed the interface problem – and in particular with solving the problem of how propositional attitudes can be causally efficacious in generating both behavior and other propositional attitudes. It is, as we shall see below, very much an open question how much of thinking behavior is even a candidate for being understood in these terms.[11]

Let me end this chapter with an overview of the key claims of the picture of the representational mind.

Checklist for the representational mind

- The representational picture makes a sharp distinction between the two components of a propositional attitude – that is, between the propositional content and the attitude taken to that content.
- The representational picture appeals to considerations of functional role to explain the attitude taken to a particular propositional content, on the assumption that there will be, for example, particular functional roles characteristic of beliefs as opposed to desires.

11 In fact, as we will see in Chapter 10, Fodor and the other proponents of the representational picture have a completely different set of arguments for the language of thought hypothesis – arguments attempting to establish the necessity of a language of thought for what we might think of as modular cognitive processes.

- An independent account has to be given of what gives a particular propositional content the content that it does.
- The representational picture holds that complexity at the level of propositional content requires an isomorphic structural complexity at the level of the vehicle.
- This requirement of structural isomorphism is met on the representational picture by taking the vehicles of propositional attitudes to be sentences in an internal language of thought.
- The causal dimension of propositional attitudes must be understood in terms of causal interactions between sentence-tokens in the internal language of thought.
- These causal interactions are a function solely of the syntactic properties of sentence-tokens in the language of thought.
- These causal transitions respect the rational relations between the contents of the attitudes in question because the language of thought is a formal system with some analog of the meta-logical property of soundness.
- This formal system is itself implemented within a physical mechanism that can operate purely causally to carry out syntactic operations.

5 Neural networks and the neurocomputational mind

- Top-down explanation vs the co-evolutionary research strategy
- Cognition, co-evolution and the brain
- Neural network models
- Neural network modeling and the co-evolutionary research paradigm

The last two chapters have considered three pictures of the mind, each offering a different way of responding to the interface problem identified in Chapter 2. The discussion so far has focused on the differences between these three pictures. From the perspective of the fourth and final picture of the mind, however, all three share certain fundamental assumptions. The framework set by those assumptions is the target of the neurocomputational picture of the mind.

The most important of these assumptions is that the direction of explanation is purely top-down, so that the way to understand the mind is to start at the top of the hierarchy of explanation and then work downwards through the levels, looking at each level for an implementation of abilities and capacities identified at the previous level. In the first section of this chapter I make this assumption explicit and outline some potential misgivings with this top-down approach. Section 5.2 explains why proponents of the neurocomputational mind propose to replace the top-down model with an investigation of cognition in which commonsense psychology co-evolves with the neuroscientific study of the brain. In section 5.3 I show how proponents of the neurocomputational picture use neural networks to pursue this co-evolutionary approach. I explain what neural networks are and how they work. Section 5.4 uses neural network models of language acquisition to explore the picture of the mind emerging from this co-evolutionary research paradigm.

5.1 Top-down explanation vs the co-evolutionary research strategy

All three pictures of the mind we have so far considered make a virtue out of abstracting away from the bottom-level details of how the brain works. The theorists we have considered up to now, be they autonomy theorists, functional theorists or computational theorists, all agree that we cannot and should not try to understand the mind by understanding the brain. Real

understanding is gained, on these views, at a level of abstraction at which we can think about cognitive tasks and how they are carried out without going into the details of the mechanisms that might actually be doing the work.

This abstraction away from the "machinery of cognition" is clearest on the picture of the autonomous mind. Autonomy theorists maintain that there is such a radical incommensurability between personal-level cognition and subpersonal mechanisms that it is very difficult to say anything informative about how they might relate to each other. The normativity and rationality at the heart of personal-level understanding of ourselves and how we behave effectively preclude any genuine vertical explanatory relations holding between personal-level horizontal explanations and subpersonal-level horizontal explanations. It may well be, as Davidson suggests, that the mental states featuring in personal-level explanations should be identified with brain states. But these identities are brute facts that neither have explanations themselves nor shed any explanatory light. This explanatory insulation between personal and subpersonal levels holds equally on Dennett's distinction between the (personal-level) intentional stance and the (subpersonal) design and physical stances. There are no interesting explanatory relations extending upwards from lower stances to the intentional stance, because the generalizations holding at the intentional stance have no echo at the lower levels. At best, an analysis at the design stance will allow us to see how a physical system can mimic the order and rationality postulated at the intentional stance. (For further discussion, see section 6.1.)

In contrast to this claim of radical incommensurability the functional and computational pictures do allow for genuine vertical explanatory relations between the different levels of explanation. But both pictures of the mind tend to view these explanatory relations in a top-down manner. Many functional and computational theorists hold that there is a privileged level of description for cognitive performances and capacities, and agree on what that privileged level of description is. It is widely held that our privileged level of description is fixed by the concepts and generalizations of commonsense psychology. Once we have a satisfactory characterization at the level of commonsense psychology, we will be able to use that characterization to pick out the physical structure that realizes the propositional attitude in question. Crudely put, we look to see what physical structure enters into the causal interactions identified by the functional analysis. We start at the top and work down.

Nor is this top-down approach the sole preserve of philosophical functionalism. Psychological functionalism takes issue with the idea that there is a simple bipartite distinction between a given functional role and the physical structure that realizes that functional role. Theorists such as Lycan (1987) suggest that there are indefinitely many different levels of description and explanation and that the distinction between functional role and realizer of that role is relative rather than absolute. Something can be a realizer relative to one level of explanation (the next level of explanation up) and at the

same time a functional role relative to another level of explanation (the next level of explanation down). The role/realizer relation extends all the way down the hierarchy of explanation. But, wherever one is in the hierarchy, it will always be the case both that we identify realizers by working downwards from the functional specification offered at the next level up and that we can fully understand a functional specification without knowing anything about its realization.

The computational picture tends also to be developed in a top-down manner. Many computational theorists adopt a distinction analogous to that between software and hardware in computer systems. The privileged level of description is algorithmic – the level that defines the particular computational task being performed and the representational primitives over which that computational task is defined. Just as a single piece of software can be run on widely divergent types of hardware, so too can the computation be carried out by very different physical systems. (Block, 1995, does his best to unpack the metaphor.)

Both the functional and the representational approaches are driven by a particular understanding of the relations that hold between scientific theories formulated at different levels of explanation, and in particular by the idea that we can, having identified a functional role at one level of explanation, proceed to identify the realizer of that role at the next level of explanation down. In the case of commonsense psychology, the functional roles are causal roles. But causal roles are only one form of functional role. The overarching category is composed of what might be termed theoretical roles. Theoretical roles are identified in terms of a particular position in the nexus of theoretical laws and principles governing a given level of explanation. We can view causal roles as special instances of theoretical roles, recognizing that in much of science the theoretical laws and principles are not causal.[1] The functional and representational approaches to the mind derive their core idea (which is, of course, a core idea about causal roles and their realizers) from an analogy with the way in which non-causal theoretical roles and their realizers are thought to work in various parts of the natural and special sciences. In motivating this analogy theorists frequently return to a small number of examples.

A clear account of the type of example on which the analogy is based can be found in the following passage from Frank Jackson, a leading exponent of the functional picture of the mind. The example comes from one of the classic examples of a scientific reduction – the reduction of thermodynamics to statistical mechanics. Here is how Jackson describes the reduction:

1 I am assuming here that differential equations plotting the relations between the rates of change of different quantities are not causal laws. For a different way of thinking about causal laws, see Woodward (2003).

> We have a story about gases told in terms of temperature, volume, and pressure; the account known as the thermodynamic theory of gases. We discover that by identifying gases with collections of widely separated, comparatively small, relatively independently moving molecules, and identifying the properties of temperature, pressure and volume with the appropriate molecular properties – temperature (in ideal gases) with mean molecular kinetic energy, for famous example – we can derive the laws of the thermodynamic theory of gases from the statistical mechanics of molecular motion, and thereby explain them (and, moreover, explain the exceptions to them) ... The whole exercise is described as a smooth reduction because the laws of the reduced theory, the thermodynamic theory of gases, are pretty much preserved in, by virtue of being pretty much isomorphic with, the corresponding laws in the reducing theory, the molecular or kinetic theory of gases.
>
> (1998, p. 57)

When we look in more detail at how the reduction works, Jackson claims, we will find that the relation between theoretical role and realizer is doing all the work.

> The discoveries that lead to the molecular theory of gases show that mean molecular kinetic energy plays the temperature role ... in the ideal gas laws. The readiness of scientists to move straight from this discovery to the identification of temperature in gases with mean molecular kinetic energy told us what their concept of temperature in gases was. It was the concept of that which plays the temperature role in thermodynamic theory of gases ... All the causal work we associate with temperature in gases is distinct from, but correlated with, mean molecular kinetic energy.
>
> (ibid., p. 58)

The picture is seductively simple. At the macro-level of observable behavior in gases we formulate general laws governing the behavior of ideal gases that relate quantities such as temperature, pressure and volume. Within the theory formed by those general laws we can identify theoretical roles corresponding to each of those concepts. We then apply those theoretical roles at the micro-level to pick out quantities at the micro-level that occupy those roles. In the case of temperature, so the story goes, the relevant quantity is mean molecular kinetic energy.

It is clear how this can provide both a model for functionalist approaches to the mind and a template for thinking more generally about the relation between levels of explanation. Each level of explanation is autonomous, consisting of laws that collectively identify a range of theoretical roles. These theoretical roles provide the links that connect each level of explanation with the level immediately below. We can start at the top with common-

sense psychology and work downwards through the levels of explanation until we arrive at levels of explanation, to do with for example the molecular biology of the neuron, where we are clearly outside the domain of the psychological.

One central motivation for the neurocomputational approach to the mind is a clear rejection of this way of thinking about the relation between levels of explanation. Supporters of the neurocomputational approach eschew top-down models of explanation and the concomitant idea that we can step neatly between autonomous levels of explanation, each of which can be understood on its own terms. This rejection is motivated in part by considerations from the philosophy of science. Supporters of the neurocomputational approach think that when we look in more detail at how scientific theories develop we find complex forms of co-evolution and interaction across levels of explanation. The relation between theories is far messier than the top-down model suggests. We should not view science as made up of a hierarchy of relatively autonomous theories that can be connected up by means of the thread of the role/realizer relation. Instead we should adopt what Patricia Churchland has called the *co-evolutionary research ideology* and take this, rather than the role/realizer model, as our guide for thinking about the relation between different levels of theorizing about the mind (P. S. Churchland 1986). But the neurocomputational approach is of course also motivated by considerations specific to the mind. Its supporters think that an account of cognition needs to be sensitive to the constraints imposed by the neural machinery in which cognition is ultimately realized. We cannot expect to find a straightforward neural implementation for theories formulated in abstraction from the concrete context within which thinking, perceiving and action take place.

Let us consider the more general thesis first. A series of studies in the philosophy of science has cast doubt on just about every aspect of the traditional view of intertheoretic reduction. It has been argued with considerable plausibility that there are no, or almost no, reductions displaying the neat structure envisaged in the quoted passage from Jackson (Feyerabend 1962; Schaffner 1967; Hooker 1981; Smith 1993). There simply does not exist the type of isomorphism between different levels of explanation required for it to be possible to identify at one level the realizer of the roles identified at another level. Even the smoothest of scientific reductions involves complicated two-way interactions between the central laws and concepts of the reduced theory and the central laws and concepts of the reducing theory.

This complexity is particularly clear even in the classical example on which Jackson and others have laid so much stress, namely, the identification of mean molecular kinetic energy as the realizer of the temperature role. Mean molecular kinetic energy is supposed to be the quantity at the microscopic level that realizes the role in generating the ideal gas laws carved out by the concept of temperature at the macroscopic level. But things are much

messier than this. There are two basic respects in which thermodynamics and statistical mechanics are fundamentally different. So fundamentally different, in fact, that it is very doubtful whether one can talk about anything at the level of statistical mechanics occupying the temperature role. It is worth looking briefly at the issues here, to get a feel for the complexities that arise when one starts to think in a little more detail about how scientific theories interface with each other.

First, the probabilistic nature of statistical mechanics has problematic consequences for some of the central concepts of thermodynamics, particularly the concept of entropy (Sklar 1993, Chapter 9; Sklar 1999; Callender 1999). At the level of statistical mechanics one has to take into account the highly improbable, but nonetheless according to the theory genuinely possible, situations in which the entropy (that is, the degree of disorder) of a system in equilibrium will spontaneously *de*crease. This is in *prima facie* conflict with the Second Law of Thermodynamics, which states, roughly, that the amount of entropy in a system will never decrease. This is important because the temperature role, however it is understood thermodynamically, must include the relation between temperature and the entropy captured in the Second Law. But it does not look as if the relation between temperature and entropy will hold in statistical mechanics in anything like the way in which it does in thermodynamics. So how then can *anything* at the level of statistical mechanics play the temperature role?

Second, thermodynamics is quintessentially time-asymmetric. That is to say, the laws of thermodynamics govern processes at the macroscopic level that run forwards in time. Indeed, the Second Law of Thermodynamics has been thought by many physicists and some philosophers to capture and explain the "arrow of time" (Price, 1996, is an accessible discussion). And yet the equations of statistical mechanics are *time-reversal invariant*. They are such that, for any state S from which a system evolves according to those equations, that system will eventually return either to S or to a state arbitrarily close to it. Once again, there is a *prima facie* problem in seeing how anything at the level of statistical mechanics can realize the temperature role. Is it not part of the temperature role that temperature should be a quantity that features in time-asymmetric processes?

So how then should we see the relation between thermodynamics and statistical mechanics? The theories are certainly not incommensurable. Nor are the central concepts of thermodynamics in any sense superseded by the concepts of statistical mechanics. What has actually happened is that the tensions between the principles of thermodynamics and the principles of statistical mechanics have had unexpected ramifications for both theories. Concepts such as the concept of entropy, which is a unitary concept in classical thermodynamics, have become fractionated in a way that has ramifications both at the level of statistical mechanics and at the level of thermodynamics (Sklar 1993, Chapter 3). Some of the flavor of the *co-evolution* of thermodynamics and statistical mechanics is conveyed in the following

passage from Clifford Hooker's pioneering study of intertheoretic reduction:

> First, the mathematical development of statistical mechanics has been heavily influenced precisely by the attempt to construct a basis for the corresponding thermodynamical properties and laws. For example, it was the discrepancies between the Boltzmann entropy and thermodynamical entropy that led to the development of the Gibbs entropies, and the attempt to match mean statistical quantities to thermodynamical equilibrium values which led to the development of ergodic theory. Conversely, thermodynamics is itself undergoing a process of enrichment through the injection "back" into it of statistical mechanical constructs, e.g. the various entropies can be injected "back" into thermodynamics, the differences among them forming a basis for the solution of the Gibbs paradox. More generally, work is now afoot to transform thermodynamics into a generally statistical theory, while retaining its traditional conceptual apparatus, and there is some hope that this may allow its proper extension to non-equilibrium processes as well.
>
> (Hooker 1981, p. 49)

Far from there being a smooth reduction of thermodynamics to statistical mechanics, the fundamental tensions between the two theories have led to each being revised and reshaped in terms of the other. There is no single direction of explanation, but rather a two-way process of interaction and accommodation.

One should not read too much into a single example, but this gives us some grounds for distrusting the idea that the role/realizer model can hold the key to thinking about the relation between different levels of explanation in studying the mind – particularly when we remember that the relation between temperature and mean molecular kinetic energy is frequently put forward as the paradigm application of the role/realizer model. If the role/realizer model does not apply even in the classic textbook examples of scientific reductions, then why should we assume in advance that it will be the Ariadne's thread leading us through the vertical relations holding between different levels of theorizing about the mind? Would it not be more sensible to think that theories in the cognitive and behavioral sciences will co-evolve, rather than fit together neatly in the top-down manner envisaged by supporters of the functional and computational approaches?

But the neurocomputational approach to the mind is not motivated simply by reflection on case histories from the philosophy of science. Interestingly enough, some of the further concerns specific to the case of cognition arise from thinking, not about the similarities between inter-level relations in the scientific study of the mind and those holding elsewhere in science, but rather about the differences. In particular, there are certain fundamental questions to be asked about the very notion of commonsense psychology, even when it is taken on its own terms and considered as an

autonomous theory. The issue is not just that it is an open question whether the categories of commonsense psychology will prove to be in any sense fundamental kinds. That question can arise for the categories of any theory. The problem is more fundamental. For supporters of the neurocomputational approach to the mind it really is an open question to what extent, even on its own terms, commonsense psychology is an explanatory theory of the sort that might define theoretical roles that can serve as links between different levels of explanation.

It is important to keep two issues separate here. Some prominent supporters of the neurocomputational approach to the mind have defended versions of radical *eliminativism*, according to which commonsense psychology is radically false (Churchland 1981). We will consider this debate in more detail in subsequent chapters. Clearly, radical eliminativism is one way of arguing that commonsense psychology will not be integrated with lower levels of explanation via the role/realizer relation. But one can have doubts about the applicability of the role/realizer relation to commonsense psychology without being an eliminativist. And in fact the arguments for eliminativism put forward by Churchland all rest on the assumption that commonsense psychology is a putative explanatory theory. Churchland considers commonsense psychology to be the sort of thing that would, if it were true, define theoretical roles that could in principle serve as a bridge to lower levels of explanation. But there are some very real questions about whether this is the right way to approach commonsense psychology. For proponents of the autonomous mind, for example, commonsense psychology is an essentially normative theory, and hence not at all the sort of thing that can be used to define theoretical roles with realizers at lower levels of explanation. Supporters of the simulationist approach to commonsense psychology, on the other hand, hold that social understanding involves simulating another person's mental states, rather than subsuming their behavior under commonsense psychological generalizations (Gordon 1986; Heal 1986). If commonsense psychological understanding is simulation-driven rather than theory-driven, then it is hard to see how commonsense psychology could be a source of theoretical roles that will provide a vertical resolution of the interface problem.

One might, in the light of this be struck by the thought that there is so little consensus about the status and nature of commonsense psychology that there are no prospects for thinking that it will, on its own terms, be capable of generating the theoretical roles required for it to be integrated with lower levels of explanation according to the realizer/role model. Commonsense psychology is simply not an agreed-upon descriptive theory of the macroscopic behavior of persons in the way that, say, classical thermodynamics is an agreed-upon descriptive theory of the behavior of heat and its transformation into mechanical energy. So, given the doubts expressed earlier about the appropriateness of the role/realizer model even in cases such as the relation between classical thermodynamics and statistical mechanics, it

might seem plausible to view the relation between commonsense psychology and the various subpersonal levels of explanation as one of co-evolution rather than as involving the identification at the various subpersonal levels of realizers for roles picked out at the level of commonsense psychology.

As far as the underlying motivation for the neurocomputational picture of the mind is concerned, stress is usually laid on the arguments from eliminative materialism. It is more profitable, however, to view eliminative materialism as just one strand within the neurocomputational approach. What really motivates the neurocomputational approach to the mind is the rejection of the top-down explanatory model associated with the functional and computational pictures in favor of a co-evolutionary conception of intertheoretic relations. There is a spectrum of different ways in which that co-evolutionary conception can be developed. Eliminative materialism, as developed by Paul and Patricia Churchland, occupies an extreme position on that spectrum. The co-evolution that they envisage between commonsense psychology and the various subpersonal levels of explanation will involve a radical revision, and ultimately (they think) a displacement of commonsense psychology. But it is nonetheless a co-evolution. It will be through thinking about the relation between the states, skills and abilities falling within the domain of commonsense psychology, on the one hand, and the various tools that neuroscience offers for understanding the brain, on the other, that we will eventually arrive at such a displacement of folk psychology. Paul Churchland's explicit arguments for eliminative materialism are best seen, not as free-standing attempts to convince us of the truth of eliminative materialism *ab initio*, but rather as predictions of the eventual results of the co-evolution of folk psychology and the various branches of neuroscience. And of course one can think that commonsense psychology will co-evolve with neuroscience and neurobiology without thinking that it will ultimately be displaced by those subpersonal theories. There is plenty of room within the broadly neurocomputational approach to the mind for less extreme ways of thinking about the future of commonsense psychology. The defining feature of the neurocomputational mind is the thought that our personal-level theories of cognition must co-evolve with our subpersonal theories of the mechanisms of cognition. One might think, given how little we actually do know about the details of how higher-level cognitive functions are realized in the brain, that a degree of agnosticism about the outcomes of the co-evolution would be a prudent strategy.

5.2 Cognition, co-evolution and the brain

A piece of the jigsaw remains missing, however. The previous section suggested (on behalf of the neurocomputational approach) that there are good reasons for thinking that commonsense psychology cannot be viewed as an autonomous level of explanation connected up with lower levels of explanation through the role/realizer relation. We should instead view it as a theory

that will co-evolve with developments in subpersonal approaches to the mind. But there are many levels of explanation between commonsense psychology and even the most abstract and high-level forms of neuroscience. What makes supporters of the neurocomputational mind so confident that these are not immediately relevant to studying the mind? Why should the general idea of co-evolution lead us to the neurosciences? Why should commonsense psychology not co-evolve with a higher-level theory such as cognitive psychology, for example?

Defenders of the neurocomputational approach confronted with this question would, I think, reply that nothing short of co-evolution between commonsense psychology and neuroscience will solve the general problems that have been identified for thinking about the role/realizer relation as a model for how scientific theories mesh together. These problems will apply with equal force if commonsense psychology is taken to co-evolve with, for example, cognitive psychology or computational psychology. Such an approach would not be an alternative to the top-down strategy. It would merely be a different way of implementing that strategy. We can see this by working through an example.

Suppose, for example, that we are working with a distinction between levels of explanation similar to that proposed by Marr and discussed in Chapter 2. Simplifying somewhat, we might locate commonsense psychology at Marr's computational level of analysis and consider the possibility of the computational and algorithmic levels co-evolving. This would mean, for example, that the specification of the information-processing task would change as a function of the range of available algorithms and the representational primitives over which they are defined – as well as being modified by what happens when the algorithms are actually run. The algorithmic level, however, would remain autonomous, completely describable without taking into account how the relevant algorithms might be implemented at the physical level.

But why, one might ask, should the algorithmic level be any more immune to bottom-up constraints from the practicalities of implementation than the computational level is immune to bottom-up constraints from the practicalities of algorithmic analysis? There seem to be plenty of reasons why one might think that the line should *not* be drawn at the algorithmic level. Is it really reasonable to think that an explanatorily adequate algorithm can be formulated without thinking about whether and how it might be implemented? Everything depends upon what the purpose of formulating the algorithm is intended to be. Since we are interested in how commonsense psychology might interface with lower levels of explanation, we are looking for an algorithm that is psychologically plausible. That is to say, we are looking for an algorithm that approximates to how the computational task in question is actually carried out by the cognitive system, rather than one that simply comes up with the same output. But this requirement of psychological plausibility imposes a range of constraints upon how the algorithm is formulated.

One set of constraints derives from the temporal dimension of cognition. Cognitive activity needs to be coordinated with behavior and adjusted on-line in response to perceptual input. The control of action and responsiveness to the environment requires cognitive systems with an exquisite sense of timing. The right answer is no use if it comes at the wrong time. Suppose, for example, that we are thinking about how to model the way the visual system solves problems of predator detection. In specifying the information-processing task we need to think about the level of accuracy required. It is clear that we will be very concerned about false negatives (i.e. thinking that something is not a predator when it is), but how concerned should we be about false positives (i.e. thinking that something is a predator when it is not)? There is a difference between a model that is designed never to deliver either false positives or false negatives and one that is designed simply to avoid false negatives. But which model do we want? It is hard to see how we could decide without experimenting with different algorithms and seeing how they cope with the appropriate temporal constraints. The ideal would be a system that minimizes both false negatives and false positives, but we need to factor in the time taken by the whole operation. It may well be that the algorithm that would reliably track predators would take too long, so that we need to make do with an algorithm that merely minimizes false negatives. But how can we calculate whether it would take too long or not? We will not be able to do this without thinking about how the algorithm might be physically implemented, since the physical implementation will be the principal determiner of the overall speed of the computation.

Supporters of the neurocomputational approach to the mind often refer in this context to the so-called 100-step rule, which has to do with the constraints upon computational speed imposed by the physical structure of neurons (Feldman and Ballard 1982). It is suggested that the minimum time in which the brain could carry out a computational task is 5 milliseconds (where a millisecond is a thousandth of a second), based on the time it takes for a neuron to generate an action potential. Many highly important computational tasks such as visual recognition, however, take no more than 500 milliseconds. This, so it is argued, imposes constraints upon the types of algorithm that can carry out these computational tasks. In particular, no such algorithm can require more than 100 computational steps.

Now, it is not clear that there is any such 100-step rule. It is not obvious, for example, that the time it takes a *neuron* to generate an action potential should determine the minimum time in which the *brain* can carry out a computational step. But the point at issue is more general. Whether or not the structure of the brain does impose a 100-step rule, the fact remains that it will inevitably impose certain time-related constraints and that these constraints will circumscribe the way in which we understand specific types of computation and information-processes at the algorithmic level. There is no sharp divide to be made between our understanding of the algorithms governing cognition and our understanding of the mechanisms in which those algorithms are realized.

We cannot expect to identify roles at the algorithmic level and then look for realizers for those roles at the implementational level.

A further way that the practical requirements of implementation impose constraints upon how we think about cognitive functioning at higher levels emerges from an aspect of cognitive abilities rarely taken into account by philosophers of mind and philosophers of psychology. We cannot consider cognitive skills and abilities from a purely synchronic perspective. The mind is not a static phenomenon. Cognitive abilities and skills themselves evolve over time, developing out of more primitive abilities and giving rise to further cognitive abilities. Eventually they deteriorate and, for many of us, gradually fade out of existence. In some unfortunate cases they are drastically altered as a result of traumatic damage. A theory of the mind needs to take into account these diachronic features of the large-scale dynamics of cognition. This imposes a range of further constraints. An account of the mind must be compatible with plausible accounts of how cognitive abilities emerge. It must be compatible with what we know about how cognitive abilities deteriorate. It must be compatible with what we know about the relation between damage to the brain and cognitive impairment.

The facts driving each of these constraints derive directly from the fact that minds are realized in brains. We know, for example, that cognitive abilities tend to *degrade gracefully*. Cognitive phenomena are not all-or-nothing phenomena. They exhibit gradual deterioration in performance over time. As we get older, reaction times increase, motor responses slow down and recall starts to become more problematic. But these abilities do not (except as a result of trauma or disease) suddenly disappear. The deterioration is gradual, incremental, and usually imperceptible within small time frames. No account of cognition can afford to ignore this, and certainly not afford to be incompatible with it. But the general phenomenon of graceful degradation is a function of the fact that the cognitive abilities involved are neurally implemented. Once again, we can expect to see a co-evolution, with details of how cognitive abilities deteriorate feeding into higher-level accounts of those abilities and those higher-level accounts feeding back into our understanding of the brain.

Similar points apply to our understanding of how cognitive abilities emerge and develop. The process of language acquisition, for example, has a characteristic dynamic profile (Barrett 1995; Elman *et al.* 1996). There are periods of rapid acceleration in vocabulary acquisition and periods of almost static consolidation. There is a clear progression in levels of syntactic and grammatical complexity, with more or less determinate stages that hold not only across individuals within a particular linguistic population, but even across different linguistic populations. No account of what it is to understand a language can be incompatible with what we know from developmental psychology and linguistics about how a language is learnt. But it is hard to see how the patterns of language learning can be understood without investigating the details of the neural processes that subserve language learning. Brains learn the way they do because of how they are constructed – and in

particular because of the patterns of connectivity existing at each level of neural organization (between neurons, populations of neurons, neural systems, neural columns, and so forth). We would expect our higher-level theories of cognitive abilities to be constrained by our understanding of the mechanisms of learning – and, of course, our understanding of the mechanisms of learning will be constrained by our account of what it is that is being learnt. So, once again, the co-evolution of theories is to be expected.

The neurocomputational approach is driven by the two basic thoughts. The first (discussed in section 5.1) is a rejection of the top-down model of explanation in favor of a co-evolutionary conception of how different levels of explanation interlock and interact. The sources for this rejection lie both in general reflections on the complexities of the relations between levels of explanation more generally within science and in concerns specific to thinking about the mind. The second is that personal-level thinking about the mind will need to co-evolve primarily with the sciences studying the neural dimension of cognition. In order to see how these motivations get translated into practice, however, we need to look in more detail at precisely how the co-evolution is supposed to work in this specific case. That will be the task of the next section.

5.3 Neural network models

An obvious obstacle to pursuing the type of co-evolutionary research strategy under discussion is the difficulty of establishing a direct interface between the categories of commonsense psychology and personal-level explanation, on the one hand, and the direct study of the brain, on the other. There are many intervening levels of explanation between commonsense psychology and neuroscience (and, in fact, within neuroscience itself). So, how can our understanding of higher-level cognition and ordinary commonsense psychology co-evolve with our understanding of the brain? Where are the points of contact that would make such a co-evolution possible?

It is true that recent years have seen an enormous increase in detailed knowledge of how the brain works. The techniques of neuro-imaging, such as *functional magnetic resonance imaging* (fMRI) and *positron emission tomography* (PET), have allowed neuroscientists to begin establishing large-scale correlations between types of cognitive functioning and specific brain areas (Posner and Raichle 1994; Buckner and Petersen 1998). These techniques work by measuring changes in blood flow, which is known to be correlated directly with cognitive functioning. PET and fMRI scans allow neuroscientists to identify the neural areas that are activated during specific tasks. These techniques make an important contribution to allowing a functional map to be built up of the brain, usefully supplementing the information available from studies of brain-damaged patients. Other techniques have made it possible to study brain activity (in non-human animals, from monkeys to sea-slugs) at the level of the single neuron (Stein *et al.* 1998). Microelectrodes can be

used to record electrical activity both inside a single neuron and in the vicinity of that neuron. Recordings from inside neurons allow a picture to be built up of the different types of input to the neuron, both excitatory and inhibitory, and of the mechanisms that modulate output signals. Extracellular recordings, on the other hand, allow researchers to track the activation levels of individual neurons over extended periods of time and to investigate how particular neurons respond to distinct types of sensory input and how they discharge when particular motor acts are performed.

Neither of these ways of studying the brain is well suited to pursuing the co-evolutionary research paradigm. The problem is one of fineness of grain. To put it crudely, the various techniques of neuro-imaging are too coarse-grained and the techniques of single neuron recordings too fine-grained (at least for studying higher cognitive functions). PET and fMRI are good sources of information about which brain areas are involved in particular cognitive tasks, but they do not tell us anything about how those cognitive tasks are actually carried out. A functional map of the brain tells us very little about how the brain actually carries out the functions in question. We need to know not just *what* particular regions of the brain do, but *how* they do it. Nor will this information come from single neuron recordings. We may well find out from single neuron recordings in monkeys that particular types of neuron in particular areas of the brain respond very selectively to a narrow range of visual stimuli, but we have as yet no idea how to work up from this to an account of how vision works.

Everything we know about the brain suggests that we will not be able to understand cognition unless we understand what goes on at the levels of organization in between large-scale brain areas and individual neurons.[2] The brain is an extraordinarily complicated set of interlocking and interconnected circuits. The most fundamental feature of the brain is its *connectivity* and the crucial question in understanding the brain is how distributed patterns of activation across populations of neurons can give rise to perception, memory, sensori-motor control and high-level cognition. But we have (as yet) limited tools for directly studying how populations of neurons work.[3]

2 This is not to say, of course, that we do not need to understand what goes on at the level of the individual neuron, or even below at the cellular and molecular levels – merely that this is unlikely to be the whole story. For a different perspective on the appropriate level for studying the brain, see Bickle (2003a).

3 There are ways of directly studying the overall activity of populations of neurons (Hillyard 1999; Bressler 2003). Event-related potentials (ERPs) and event-related magnetic fields (ERFs) are cortical signals that reflect neural network activity and that can be recorded in a non-invasive manner from outside the skull. Recordings of ERPs and ERFs have the advantage over information derived from PET and fMRI of permitting far greater temporal resolution and hence of giving a much more precise sense of the time course of neural events. Yet information from ERPs and ERFs is still insufficiently fine-grained. They reflect the summed field potentials of populations of neurons, but offer no insight into how that summed field potential emerges from the activity of individual neurons.

Using microelectrodes to study individual neurons can provide no clues to the complex patterns of interconnection between neurons. Single neuron recordings will tell us what the results of those interconnections are for the individual neuron, as they are manifested in action potentials, synaptic potentials and the flow of neurotransmitters, but not about how the behavior of the population as a whole is a function of the activity in individual neurons and the connections between them. At the other end of the spectrum, large-scale information about blood flow in the brain will tell us which brain systems are active, but is silent about how the activity of the brain system is a function of the activity of the various neural circuits of which it is composed.

It is for this reason that a powerful tool in the neurocomputational approach to the mind is a form of modeling, rather than direct empirical study. Since we do not have the equipment and resources to study populations of neurons directly, many researchers have thought that the most promising strategy is to develop models that approximate in certain important respects to populations of neurons and investigate how they can carry out different types of cognitive tasks. We can distinguish two different types of explanatory project in this area (Arbib 2003, p. 3). The emphasis in *computational neuroscience* is on modeling biological neurons and populations of biological neurons, whereas the project of *neural computing* abstracts away much more from biological details in the interests of computational tractability and technological utility. What I am terming the neurocomputational approach to the mind can be developed in either of these directions (and in fact the difference between them is really only one of degree). However, since much computational neuroscience and neural computing is formidably complex, I will be focusing on the class of models that has received the most philosophical attention and that can be most easily understood by non-mathematicians. (See the annotated bibliography for references to other types of model.)

These models, sometimes called *connectionist* networks and sometimes *artificial neural networks*, are mathematical models that make use of the powerful resources of modern digital computers (Bechtel and Abrahamsen 1991; Churchland and Sejnowski 1992; McLeod *et al.* 1998). Artificial neural networks abstract away from many biological details of neural functioning in the hope of capturing some of the crucial general principles governing the way the brain works. They aim to reduce the multilayered complexity of brain activity to a relatively small number of variables whose activity and interaction can be rigorously controlled and studied. As one might expect, there are many trade-offs in neural network modeling between computational tractability and biological plausibility, and there are very real questions to be asked about the fit between artificial neural networks and the genuine neural networks that they model. We will return to those questions later in the chapter. In the remainder of this section I will outline some of the general characteristics of neural network models and the most general

structural parallels between artificial neural networks and the brain circuits that they model.

It will be useful to begin by outlining some of the key features of artificial neural networks. The first is that they involve parallel processing. An artificial neural network contains a large number of units (which might be thought of as artificial neurons). Each unit has a varying level of activation, typically represented by a real number between -1 and 1. The units are organized into layers with the activation value of a given layer determined by the activation values of all the individual units. The simultaneous activation of these units, and the consequent spread of activation through the layers of the network, govern how information is processed within the network. The second key feature is that each unit in a given layer has connections running to it from units in the previous layer (unless it is a unit in the input layer) and will have connections running forward to units in the next layer (unless it is a unit in the output layer).[4] The pattern of connections running to and from a given unit is what identifies that unit within the network. The strength of the connections (the *weight* of the connection) between individual neurons varies and is modifiable through learning. This means that there can be several distinct neural networks each computing a different function, even though each is composed of the same number of units organized into the same set of layers and with the same connections holding between those units. What distinguishes one network from another is the pattern of weights holding between units. The third key feature is that there are no intrinsic differences between one unit and another. The differences lie in the connections holding between that unit and other units. The fourth feature of most artificial neural networks is that they are trained, rather than programmed. They are generally constructed with broad, general-purpose learning algorithms that work by changing the connection weights between units in a way that eventually yields the desired outputs for the appropriate inputs.

Let us look at how an artificial neural network is set up in a little more detail. Figure 5.1 is a schematic diagram of a generic neural network with three layers of units. The basic architecture of the network is clearly illustrated in Figure 5.1. The network is composed of a set of processing units organized into three different layers. The first layer is made up of input units, which receive inputs from sources outside the network. The third layer is made up of output units, which send signals outside the network. The middle layer is composed of what are called hidden units. Hidden units are distinctive in virtue of communicating only with units within the network. The hidden units are the key to the computational power of artificial neural networks. Networks without hidden units are only capable of carrying out a limited variety of computational tasks. The illustrated network

4 See footnote 5.

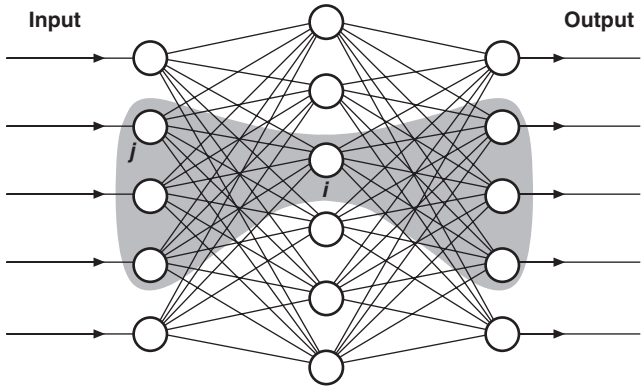

Figure 5.1 The computational operation performed by a unit in a connectionist model: the general structure of a connectionist network (source: McLeod *et al.* (1998, p. 16)).

only has one layer of hidden units, but of course networks can be constructed with as many layers as required.

The lines in Figure 5.1 illustrate the connections holding between units in the network. Two points are worth noting about these connections. The first is that there are no connections holding between units within a given layer. Hidden units are connected to input units and to output units, but not to other hidden units. The second is that all the connections hold in a single direction, forwards through the network from input to output. The network is a *feedforward* network.[5]

But how does the network actually work? How does it process information? The basic idea is that information takes the form of activation spreading through the network. Each unit within the network has an activation level that is a function of the activation levels of the units that feed into it. The end result is that a particular pattern of activation across the input units leads eventually to a particular pattern of activation across the output units. We can break this process down into stages. Input units have activation values representing features external to the network. We can think of these activation values as in some sense corresponding to the firing rates of individual neurons, although within the network they will take numerical values (typically taking some real value in the interval $[-1, 1]$). The input

5 Not all networks are feedforward networks. Recurrent neural networks have feedback connections built into them. For a brief introduction to recurrent networks, see Chapter 7 of McLeod *et al.* (1998) and pp. 115–125 of Churchland and Sejnowski (1992). Andy Clark (2001a, Ch. 4) offers a useful overview of the principal types of connectionist network currently under investigation.

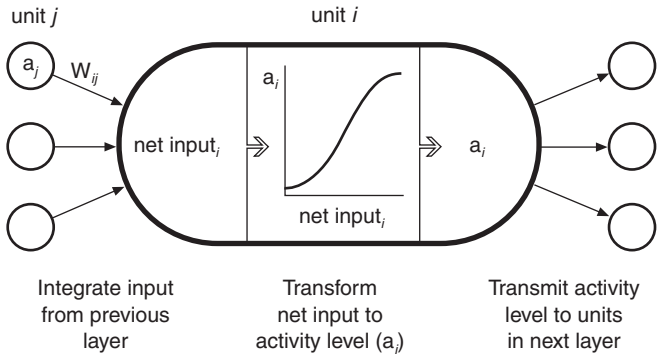

Figure 5.2 Operation of unit *i* from Figure 5.1: (1) Integrate the inputs from the previous layer to create a net input; (2) Use an activation function to convert the net input to an activity level; (3) Output the activity level as input to units in the next layer (source: McLeod *et al.* (1998, p. 16)).

to the network will consist of a pattern of activation values across the input units. In mathematical terms we can think of this as a vector (that is, an ordered set of numbers). The result of the processing in the network is a pattern of activation values across the output units. This pattern can also been seen as another vector. So, the information processing within the network can be viewed as the transformation of one vector into another. How does this transformation work? How does a pattern of activation values at the input layer get transformed into a pattern of activation values at the output layer?

The spread of activation through an artificial neural network is governed by the principle that the activation value of each unit is transmitted forwards through the network to each unit to which it is connected. Different connections have different "strengths". We can think of connections as being either reinforcing or inhibiting. This is reflected in the mathematics of the network through the weights that are attached to the connections between units. These weights can be either positive (increasing the activation value of the sending unit) or negative (decreasing the activation value of the sending unit). The activation value transmitted through the connection to the receiving unit is equal to the product of the weight and the activation value of the sending unit. Each unit that receives input from other units has its own activation value determined as a function of the total input it receives. This process is illustrated in Figure 5.2, which shows part of what is happening at the central hidden unit in the generic network we have been discussing.

The first step in the process is to calculate the net input to the hidden unit. The net input is the sum of the activation values of all the units from which the hidden unit receives input. The next step is to calculate the acti-

vation value of the hidden unit itself – and hence to calculate what the output of that unit will be to the other units to which it is connected. In some simple networks the activation value just is the net input, or some linear function of it, but it is more usual to build into the system some sort of threshold, so that there is no output from a processing unit until the net input reaches a certain activation level. There are different types of threshold function, varying from a simple on–off step function according to which the unit is either off or maximally activated to the somewhat more complicated sigmoid function illustrated in Figure 5.2. The sigmoid function allows the rate of increase of the activation value to vary depending on the extent to which the net input approaches its maximum possible value. Whatever activation function is chosen, the result is an activation value for the processing unit. If the processing unit is a hidden unit, then this activation value will itself be weighted and then form part of the net input into the further units to which that unit is connected – at which point the activation function will once again come into play to determine activation value as a function of net input. And so on, until the output level is reached.

What we have seen so far is how activation spreads through the network, given a particular input vector and a particular set of weights. But where do these weights come from? How can they be modified in a way that will allow the network to learn? The majority of artificial neural networks are *supervised networks*, which means that the difference between the actual output produced by the network in response to a given input and the desired output for that is used to "train" the network. Most of the training methods for artificial neural networks follow a common pattern. The details of the individual training methods are formidably complex, but the basic idea is straightforward. I will sketch the basic principles behind one of the most commonly used training methods, the method of *backpropagation of error* (also known as the *generalized delta rule*). The basic idea, as with all forms of supervised learning, is that the network "learns" by reducing its degree of error to bring the actual output closer to the desired output. A lengthy series of small incremental reductions in error eventually brings it to the desired output. Each individual reduction in the degree of error is achieved by modifying the strengths of the weights holding between units as a function of the degree of divergence between the actual output and the desired output.

In a network with just two layers of units it is easy to see how the weights can be modified to reduce the overall error. We start with the error at the output units. If we only have one layer of output units and one layer of input units, then the degree of error of each output unit will be obvious. The desired output will be a particular vector of activation values over the output units, and the actual output will be a different vector of activation values. So subtracting the one vector from the other will produce a further vector that gives the degree of error for each output unit. Once we know the error for each output unit, then it is fairly clear how to diminish it. Suppose that the

level of activation of one of the output units is too low, relative to the degree of activation of the input units. In order to decrease the error we need to increase the activation of that output unit relative to the degree of activation of the input units to which it is connected. And we can do this by increasing the weights of the positive connections leading to the output unit – and decreasing the strength of the negative connections leading to it. Similarly, if the activation level of the output unit is too high, then the way to reduce the error is to decrease the strength of the positive weights and increase the strength of the negative connections. It is straightforward to devise an algorithm that will compute the degree of error and make the corresponding weight adjustments (Bechtel and Abrahamsen 1991, pp. 71–85).

However, there is a serious difficulty in applying this training method to the majority of artificial neural networks. This method of *error gradient descent learning* relies on the degree of error of each individual unit being known – which of course requires knowing the target activation level for each individual unit. Without this we will not know how to modify the weights within the network. This means that the training method is inapplicable to networks with hidden units. In a network with hidden units it is possible to compare the actual activation levels of the output units with the target activation levels for those units, but there will be nothing against which to compare the activation levels of the hidden units. So how is the network to calculate how to change the strengths of the weights, given that all of the weighted connections will involve at least one hidden unit?

The backpropagation algorithm allows the degree of error in the activation level of a hidden unit to be calculated without there being a determinate target activation level for that hidden unit. The methodological assumption is that each hidden unit connected to an output unit bears a degree of "responsibility" for the error of that output unit. If, for example, the activation level of an output unit is too low, then this can only be because insufficient activation has spread from the hidden units to which it is connected. This gives us a way of assigning error to each hidden unit. In essence, the error level of a hidden unit is a function of its degree of responsibility for the error of the output unit to which it is connected. Once this degree of responsibility, and consequent error level, is assigned to a hidden unit, it then becomes possible to modify the weights between that unit and the output unit to decrease the error. This method can be applied to as many levels of hidden units as there are in the network. We begin with the error levels of the output units and then assign error levels to the first layer of hidden units. This allows the network both to modify the weights between the first layer of hidden units and the output units and to assign error levels to the next layer of hidden units. And so the error is *propagated* back down through the network until the input layer is reached. The details of the equations by which this is achieved are too complex to go into here (an accessible account will be found in Chapter 3 of Bechtel and Abrahamsen 1991 and the full details in Rumelhart *et al.* 1986). The important point is

that activation and error are propagated through the network in fundamentally different directions. Activation spreads forwards through the network (or at least through *feedforward* networks), while error is propagated backwards.

The process of training a network is somewhat lengthy. It is usual to begin with a random assignation of weights and then present the network with a series of training input patterns of activation, each of which is associated with a target output pattern of activation. The patterns are presented, and the weights modified by means of the backpropagation learning algorithm until errors have diminished almost to zero. This results in a distinctive and stable pattern of weights across the network. The overall success of a network can be calculated by its ability to produce the correct response to inputs on which it has not been trained. It will be useful to work through a relatively straightforward example to illustrate the sort of task that a network can be trained to do and how it proceeds. Artificial neural networks are particularly suited for pattern recognition tasks. One such pattern recognition task has become a classic of artificial neural network design. Consider the task of identifying whether a particular underwater sonar echo comes from a submerged mine, or from a rock. There are discriminable differences between the sonar echoes of mines and rocks, but there are equally discriminable differences between the sonar echoes from different parts of a single mine, or from different parts of a single rock. It is no easy matter to identify reliably whether a sonar echo comes from a mine or from a rock. Human sonar operators can do so reasonably well (after a considerable amount of practice and training), but it turns out that artificial neural networks can perform significantly better than humans (Gorman and Sejnowski 1988).

The first problem in devising a network is finding a way of coding the external stimulus as a pattern of activation values. The external stimuli are sonar echoes from similarly shaped and sized objects known to be either mines or rocks. In order to "transform" these sonar echoes into a representational format suitable for processing by the network, the sonar echoes are run through a spectral analyzer that registers their energy levels at a range of different frequencies. This process gives each sonar echo a unique "fingerprint" to serve as input to the network. Each input unit is dedicated to a different frequency and its activation level for a given sonar echo is a function of the level of energy in the relevant sonar echo at that frequency. This allows the vector of activation values defined over the input units to reflect the unique fingerprint of each sonar echo. In the network developed by Gorman and Sejnowski there are 60 input units, corresponding to the 60 different frequencies at which energy sampling was carried out. The network has one layer of hidden units. Since the job of the unit is to classify inputs into two groups, the network contains two output units – in effect, a rock unit and a mine unit. The aim of the network is to deliver an output activation vector of < 1,0 > in response to the energy profile of a rock and < 0, 1 > in response to the energy profile of a mine.

The mine detector network is a standard feedforward network (which means that activation is only ever spread forward through the network) and is trained with the backpropagation learning algorithm. Although the network receives information during the training phase about the accuracy of its outputs, it has no memory of what happened in early sessions. Or rather, more accurately, the only traces of what happened in earlier training sessions exist in the particular patterns of weights holding across the network. Each time the network comes up with a wrong output (a pattern of $< 0.83, 0.2 >$ rather than $< 1, 0 >$, for example, in response to a rock profile) the error is propagated backwards through the network and the weights adjusted to reduce the error. Eventually the error at the output units diminishes to a point where the network can generalize to new activation patterns with a 90 percent level of accuracy.

The mine–rock detection task is a paradigm of the sort of task for which neural networks are best known and most frequently designed. The essence of a neural network is pattern recognition. But many different types of cognitive ability count as forms of pattern recognition (far more than one might initially think, according to proponents of the neurocomputational mind) and the tools provided by artificial neural networks have been used to model a range of cognitive processes, such as visual recognition, chromatic perception, language acquisition, concept learning and decision-making – as well as many phenomena that are not cognitive at all (such as patterns in the movements of prices on the stock markets, the values of bonds and fluctuations in demand for commodities).[6] The richness and power of neural networks are not in any sense in doubt. But the important question is what the availability (and predictive adequacy) of these sorts of models can tell us about how the mind works.

In the previous section I suggested that we should look at artificial neural networks as ways of implementing the co-evolutionary approach to the interface problem characteristic of the neurocomputational picture of the mind. Neural networks are the bridge between personal-level analyses of cognitive abilities and the direct study of the brain. What neural networks do is allow some of what we know about how the brain works to feed into our higher-level thinking about cognitive abilities. This comes about because the models that we construct using neural network tools can force us to rethink much of what we took for granted both about the nature of the cognitive abilities at stake and about the mechanisms by which they might be carried out in the brain. Now that we have some understanding of how artificial neural networks actually work, we can proceed to look in more detail at a concrete example of how neural network models can promote the co-evolutionary research ideology. This will be the task of the next section.

6 There is a useful list of applications of neural network technology in the social sciences in Garson (1998, pp. 17–21).

5.4 Neural network modeling and the co-evolutionary research paradigm: the example of language

As we saw in the previous section, proponents of the neurocomputational picture of the mind face a fundamental problem. Rejection of the top-down model of explanation leads to the thought that our higher-level understanding of cognition must both inform and be informed by our understanding of the actual neural mechanisms of cognition. But, on the other hand, we are not yet (and perhaps never will be) in a position to study the brain at what appears to be the appropriate level of organization in order to gain a direct insight into the mechanics of cognition. For that we need to work at the medium scale, studying the behavior of populations of neurons, while the tools that we currently have at our disposal are suitable either for studying large-scale neural structures or for studying individual neurons. One solution to the problem is to use artificial neural networks to allow some aspects of what we know about neural functioning to be brought to bear on our understanding of higher-level cognitive abilities while abstracting away from much of the noise, detail and complexity attendant upon studying the brain directly. In this section I will try to illustrate how this type of approach can suggest fundamentally different ways of thinking about higher-level cognitive abilities. We will work through a single example, the example of language acquisition and mastery, to see how artificial neural network modeling offers a powerful challenge to some deeply entrenched views about the nature of representation.

This way of implementing the co-evolutionary research methodology remains a hostage to fortune in one very important sense. It will only work if artificial neural networks do indeed turn out to be a good guide as to how the brain actually works and this is a matter of some controversy. Critics of artificial neural networks often point to some of the striking dissimilarities at many different levels between neural networks and the brain. The units in artificial neural networks are often presented as being neuron-like, with the outputs from individual units described as axons and the weighted connections to other neurons described as synaptic connections (e.g. P. M. Churchland 1992, pp. 32–33). But it is important not to exaggerate the similarity. Neural network units are all homogenous, for example, whereas there are many different types of neuron in the brain – twelve different types in the neocortex alone. Artificial neural networks depend upon the possibility of any given unit sending either excitatory or inhibitory impulses, but no neurons in the mammalian brain seem to have this property. Moreover, brains are quite simply not as massively parallel as the majority of artificial neural networks. It appears that each cortical neuron is connected to a roughly constant number of neurons (approximately 3 percent of the neurons in the surrounding square millimeter of cortex). There are also questions to be asked about the relative scale of connectionist networks. The cortical column is an important level of neural organization. Each cortical

column consists of a population of highly interconnected neurons with similar response properties. A single cortical column cuts vertically across a range of horizontal layers (*laminae*) and can contain as many as 200,000 neurons – whereas even the most complicated artificial neural networks rarely have more than 5,000 units. One would expect this "scaling up" from artificial neural networks to cortical columns to bring a range of further disanalogies in its wake. In particular, genuine neural systems will work on data that are far less circumscribed than the inputs to artificial neural networks.

The most significant disanalogies, however, arise with the type of learning of which artificial neural networks are capable. Some of these are practical. As we have seen, artificial neural networks learn by modifying connection weights and even in relatively simple networks this requires hundreds and thousands of training cycles. It is not clear how much weight to attach to this. After all, the principal reason why training a network takes so long is that networks tend to start with a random assignment of weights and this is not something one would expect to find in a well-designed brain. It is true, however, that neurons are responsive to hormones and other chemicals in the environment, and that neurons can make/break connections with other neurons. Still more significant are the problems posed by the training methods for artificial neural networks. There is no evidence that anything like the backpropagation of error takes place in the brain. Researchers have failed to find any neural connections that yield error feedback to alter connection weights. Moreover, most neural networks are supervised networks and only learn because they are given detailed information about the extent of the error at each output unit. But very little biological learning seems to involve this sort of detailed feedback. Language learning is a case in point. Theorists in cognitive science and linguistics hold many aspects of language mastery, particularly its syntactic dimension, to be innate. The principal arguments for innateness hypotheses are known as arguments from the poverty of the stimulus (see, for example, Pinker 1994). Such arguments claim that there is insufficient information and feedback available for young children to learn the syntax of a language. What we think of as language learning should instead be viewed as a process of setting parameters in an innately specified language module. What gives arguments from the poverty of the stimulus their power is that the feedback in learning is typically diffuse and relatively unfocused – a long way away from the precise calibration of degree of error required to train artificial neural networks.

It is important to realize, however, that artificial neural networks are mathematical models, and as such they have to abstract to a certain extent from the details of neural implementation. The point has been well put by Churchland and Sejnowski discussing a model of the oculomotor system:

> From the perspective of understanding the oculomotor system, backpropagation is a tool to create a kind of wind tunnel of the nervous system,

> wherein experiments relevant to the natural state can be made and variables otherwise beyond control may be brought under control.
>
> (1992, p. 378)

The question of whether a given artificial neural network is biologically plausible needs to be considered in the context of whether it is a good model. And the mathematical models yielded by artificial neural networks are to be judged by the same criteria as any other mathematical models. The first set of criteria is predictive. The results of the network need to mesh reasonably closely with what is known about the large-scale behavior of the cognitive ability being modeled. So, for example, if what is being modeled is the ability to master some linguistic rule (such as the rule governing the formation of the past tense), one would expect a good model to display a learning profile similar to that generally seen in the average language-learner. Equally, one would expect the outputs of a model of a cognitive ability to change when the network is damaged in a way that matches the performance of normal subjects when the neural system subserving that cognitive ability is damaged. The second set of criteria is vaguer. A model needs to be designed in a manner that reflects the general structural characteristics of the phenomenon or mechanism being modeled. The model does not have to be faithful in detail to what is being modeled (for otherwise anything would be its own best model), but it does need to reflect its general principles of design and operation.

Proponents of artificial neural networks can make a fairly strong case that artificial neural networks have promise on the first of these two sets of criteria. We will see shortly how neural networks can replicate some features of the characteristic learning patterns of young children for various aspects of language learning. And many studies have generated robust correlations between the performance of neuropsychological patients and the results of "damaging" an artificial neural network, either by removing connections, by altering the thresholds of the activation functions or by randomly changing the values of some weights (collectively known as "lesioning" the network). These correlations have generated hypotheses about the underlying explanations for some of the breakdown patterns found in neuropsychological patients. Hinton and Shallice (1991), for example, used the results of lesioning an artificial neural network to explain the pattern of pronunciation errors made by deep dyslexics – in particular their tendency when asked to read a work to utter a word that bears no phonological relation to the target word but is related to it semantically ('bridge' for 'river', for example).[7] Similarly, Farah and McCelland (1991) have offered a model that, when

7 The design of the network and the details and results of the lesioning process are described in Ch. 9 of McLeod *et al.* (1998), which provides a useful summary of connectionist research into language acquisition. See also MacWhinney (2003).

lesioned, reproduces one striking feature of breakdowns in the semantic memory system, namely, that they can be restricted to particular categories and to particular sensory modalities. In both cases the organization and architecture of the model have suggested substantial revisions of existing neuropsychological models of the respective systems.

Matters are far less straightforward with the second set of criteria, those to do with the degree of match between the general principles of the model and the general structural characteristics of the phenomenon or mechanism being modeled. Clearly, it will be harder to determine whether these criteria are satisfied by any given model – one theorist's inessential detail may be another theorist's fundamental general principle. But it seems clear that there are significant ways in which the general principles governing the design of neural network models do reflect certain basic facts about brain design and the form of neural information processing. Information processing in neural networks, for example, is parallel and distributed. Information is inputted and transformed in the form of vectors of activation values, rather than discrete independently identifiable representations. This reflects what little is known of how information processing might work in the brain. So too does the highly connected nature of neural networks. Of course, the point is well taken by neural network modelers that brains are not as highly connected as the average neural network tends to be, but global connectivity is not an essential feature of artificial neural networks – and in fact network designers have found that neural network models can be made more accurate by reducing the connectivity (see Dawson 1998, Chapter 7, for this and other examples of biologically inspired modifications to neural network design – an example of co-evolution in action). It remains true that the backpropagation learning algorithm has little or no biological plausibility, but proponents of artificial neural networks will suggest that what is important is not the specific learning algorithm used, but rather the general idea that learning takes place by means by changes in the weights of connections. Moreover, there exist other learning algorithms (such as those used in competitive networks) that are more biologically plausible than backpropagation.[8] Once again, backpropagation is not the essence of artificial neural networks, but merely one way of implementing the general principles that underlie them.

Let us grant, then, that artificial neural networks are sufficiently biologically plausible to serve as bridges between explanation at the personal level and explanation at the neuronal level. The next step is to investigate how artificial neural networks might be used to implement what we have termed

8 Competitive networks employ a form of unsupervised learning in which output units compete when presented with an input until the most highly activated neuron is the only one left active. The learning rule is local and hence there is no need for information about error to be spread through the network. See Chapter 6 of McLeod *et al.* (1998).

the co-evolutionary research ideology. We can start off with a single example and then use that example to look at some more general features of co-evolution and the neurocomputational approach to the mind.

An enormous amount of research in philosophy, psychology and linguistics has been devoted to the question of what it is to understand a language. This is partly a matter of explaining what it is to understand the meaning of words (semantics) – and partly a matter of explaining what it is to understand the principles by which words are combined into sentences (syntax). The majority of those who have considered these questions have adopted a common approach. They have taken the essence of language to consist in linguistic rules, so that, for example, to understand the meaning of a word is to be in command of the rule that governs its application (e.g. the rule that the word 'dog' applies to dogs and only to dogs). There are, of course, considerable divergences about how exactly command of a rule should be understood and, correlatively, of what it is to follow a rule. These divergences have been particularly prominent in philosophical discussion of linguistic meaning and rule following.[9] We can, for present purposes, abstract away from these debates. All we need for the moment is the very general idea that understanding a language is a matter of mastering linguistic rules, both semantic and syntactic.

As many theorists have pointed out, the question of what it is to understand a language is closely connected to the question of how languages are learnt, and the rule-based conception of linguistic understanding provides a clear model of language acquisition. On this conception, the process of acquiring a language is a lengthy process of mastering the appropriate rules, starting with the simplest rules governing the meaning of everyday words, moving on to the simpler syntactic rules governing the formation of sentences and then finally arriving at complex syntactic rules such as those allowing sentences to be embedded within further sentences and those governing complex forms of anaphoric reference and the resolution of scope ambiguities. Jerry Fodor has used a version of this conception of language acquisition as an argument for the existence of an innate language of thought (Fodor 1975).[10] His argument starts off from a particular conception of the rules governing linguistic meaning. He thinks that these rules take the form of truth-rules, where truth-rules are rules specifying the referents of proper names and the extension of predicates. An example of a truth-rule for a predicate would be: '*a* is *F*' is true iff *b* is *G* (where '*a*' refers to *b* and '*G*' is deemed to be co-extensive with '*F*'). How, Fodor asks, can we learn such rules? Only, he thinks, by a process of hypothesis formation and

9 See, for example, the extensive debate provoked by Wittgenstein's analysis of rule following in the *Philosophical Investigations* and pursued in Kripke (1982). Many central essays in the debate are collected in Miller and Wright (2002).

10 This argument is discussed in more detail in section 10.6.

testing. Learning a language is a matter of forming hypotheses about what the truth-rule associated with a word might be, testing those hypotheses against further linguistic data and then making any necessary adjustments. This requires a language of thought, he argues, because hypotheses about truth-rules must be formulated in a language that cannot of course be the language being learnt. Here we have a substantive claim about the mechanics of cognition based upon a particular personal-level characterization of the process of language acquisition. It is, moreover, a characterization that can easily be extended to the syntactic dimension of language learning, so that the process of coming to understand the syntactic structure of a language is understood as a process of forming hypotheses about the syntactic rules governing how words can be combined and sentences formed.

What reasons are there for thinking that this is the best way to think about language learning and language mastery? In one sense, of course, it might seem obvious that learning a language is a matter of mastering rules. Since the syntax and semantics of a language can be specified as a set of rules, it seems natural to describe the process of learning a language as a process of coming to master those rules. But *this* falls far short of Fodor's proposal. It by no means follows that learning a language involves formulating and testing hypotheses about what these rules are. There are all sorts of ways in which one's mastery of a linguistic rule might be implicit rather than explicit, so that one learns to follow the rule without formulating a series of increasingly refined versions of it. Need there be anything more to mastering a rule than using words in accordance with that rule? Must the ability to use words in accordance with a rule be understood as a matter of in some sense *internalizing* the rule? This is, of course, an area of serious philosophical disagreement, with a broad spectrum of possible positions (with Fodor at one end and Horwich (1998), at the other – see section 10.6 for further discussion). In a sense, however, the issue is not *purely* philosophical. What is at stake is the process of language learning and one might think that there is an empirical fact of the matter about the form that this process takes. It is natural, then, to wonder whether there might be any relevant empirical evidence. Are there any facts about how languages are learnt that could point us towards one end of this spectrum, either towards the hypothesis-testing end or towards the meaning-as-use end?

Any account of language learning will have to explain how children (and, to a lesser extent, adults learning a second language) resolve the difficulties posed by the fact that languages have both regular and irregular verbs. We can take the particular and well-studied example of the past tense in English. As any non-native English speaker will know, this is a veritable minefield. There are robust data indicating that children go through three principal stages in learning how to use the past tense in English (Rumelhart and McClelland 1986, reporting data presented in Brown 1973 and Kuczaj 1977). In the first stage young language-learners employ a small number of very common words in the past tense (such as 'got', 'gave', 'went', 'was',

etc.). Most of these verbs are irregular and the standard assumption is that children learn these past tenses by rote. In the second stage children use a much greater number of verbs in the past tense, some of which are irregular but most of which employ the regular past tense ending of '– ed' added to the root of the verb. During this stage they can generate a past tense for an invented word by adding '– ed' to its root and, interestingly, make mistakes on the past tense of the irregular verbs that they had previously given correctly (saying, for example, 'gived' where they had previously said 'gave'). These errors are known as *over-regularization* errors. In the third stage children cease to make these over-regularization errors and regain their earlier performance on the common irregular verbs while at the same time improving their command of regular verbs.

It is easy to see how this pattern of performance might be thought to support something like Fodor's rule-governed conception of language learning. It might be suggested, for example, that what happens in the second stage is that children make a general hypothesis to the effect that all verbs can be put in the past tense by adding the suffix '– ed' to the root. This hypothesis overrides the irregular past tense forms learnt earlier by rote and produces the documented over-regularization errors. In the transition to the second stage, the general hypothesis is refined as children learn that there are verbs to which it does not apply and, correspondingly, begin to learn the specific rules associated with each of these irregular verbs.

What might count as evidence *against* this interpretation of the different stages in children's performance in learning the past tense? This is where artificial neural networks come back into the picture, because researchers in neural network design have devoted considerable attention to designing networks that reproduce the characteristic pattern of errors in past tense acquisition without having programmed into them any explicit rules about how to form the past tense of verbs, whether regular or irregular. The pioneering network in this area was designed by Rumelhart and McClelland (1986). It was a relatively simple network, without any hidden units (and hence not requiring backpropagation), but nonetheless succeeded in reproducing significant aspects of the learning profile of young children. The network was initially trained on ten high-frequency verbs, to simulate the first stage in past tense acquisition, and then subsequently on 410 medium frequency verbs (of which 80 percent were regular).[11] At the end of the training the network was almost errorless on the 420 training verbs and generalized quite successfully to a further set of 86 low-frequency verbs that it had not

11 To get a sense of the amount of training required for an artificial neural network, the initial training involved 10 cycles with each verb being presented once in each cycle. The subsequent training involved 190 cycles, with each of the 420 verbs (the 410 medium-frequency verbs together with the 10 original high-frequency verbs) with each cycle once again involving a single presentation of each verb.

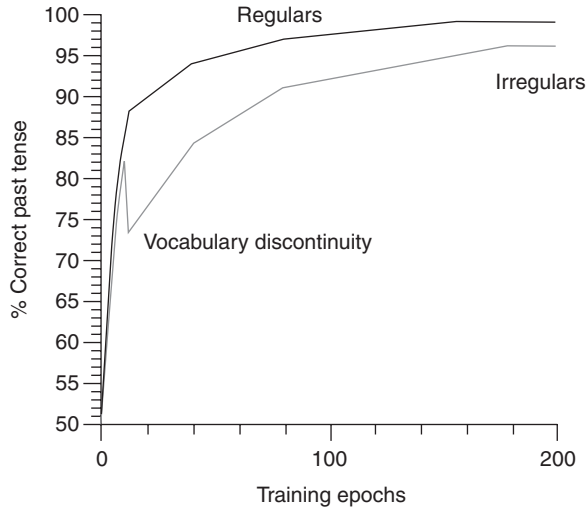

Figure 5.3 Performance on regular and irregular verbs in the Rumelhart and McClelland (1986) model of the acquisition of the English past tense. The vocabulary discontinuity at the tenth training epoch indicates the onset of overregularization errors in the network.

previously encountered (although, as one might expect, the network performed better on novel regular verbs than on novel irregular verbs).

One significant feature of the Rumelhart and McClelland network is that it reproduced the over-regularization phenomenon. This is shown in Figure 5.3, which maps the network's relative success on regular and irregular verbs. As Figure 5.3 shows, the network starts out rapidly learning both the regular and the irregular past tense forms. There is a sharp fall in performance on irregular verbs after the 11th training cycle, while the degree of success on regular verbs continues to increase. While the network's performance on irregular verbs is "catching up" with its performance on regular verbs, the characteristic errors involve treating irregular verbs as if they were regular.

As has frequently been pointed out (Pinker and Prince 1988; Prince and Pinker 1988), there are methodological problems with the Rumelhart and McClelland network. In particular, the over-regularization effect seems to be built into the network by the rapid expansion of the training set after the 10th cycle – and in particular by the fact that the expanded training set is predominantly composed of regular verbs. Nonetheless, a series of further studies have achieved similar results to Rumelhart and McClelland with less question-begging assumptions.[12] Plunkett and Marchman, for example, have

12 For more details, see Chapter 9 of McLeod *et al.* (1998) and Chapter 3 of Elman *et al.* (1996).

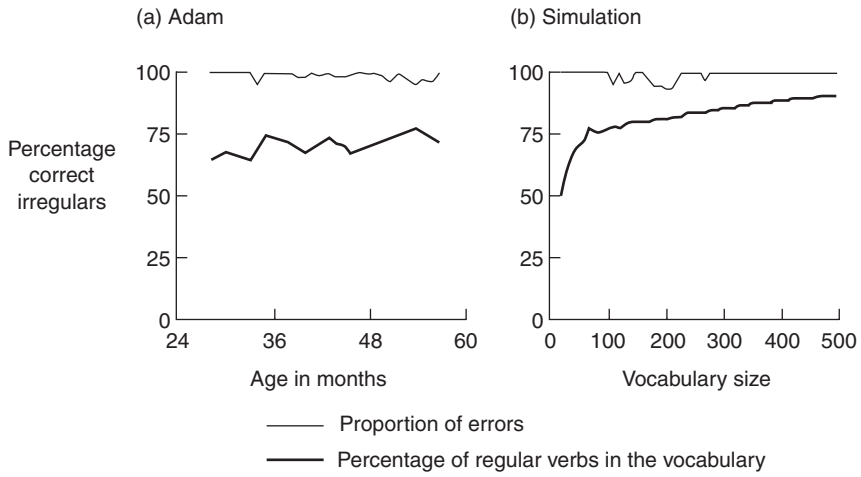

Figure 5.4 A comparison of the over-regularization errors of Adam, a child studied by Marcus *et al*. (1992) and those produced by the Plunkett and Marchman (1993) simulation. The thin lines show the proportion of errors as a function of age (Adam) or vocabulary size (simulation). The thick lines indicate the percentage of regular verbs in the child's/network's vocabulary at various points in learning (source: McLeod *et al*. (1998, p. 186)).

produced a network with one layer of hidden units that generates a close match with the learning patterns of young children. Unlike the McClelland and Rumelhart model, the vocabulary size was gradually increased and the percentage of regular verbs in the total vocabulary was 90 percent, which matches more or less the relative frequency of regular verbs in English. It is interesting to compare the learning profile of the Plunkett and Marchman network with the detailed profile of the learning pattern of a child studied by Marcus *et al.* (1992). Figure 5.4 compares the percentage of correctly produced irregular past tenses in the Plunkett and Marchman simulation and in a child whose past tense acquisition was studied by Marcus and colleagues.

As Figure 5.4 shows, the percentage of correctly produced irregular past tenses drops in both the network and the child as the vocabulary size increases. This might be thought to correspond to the second of the three stages identified earlier and to be correlated with the predominance of over-regularization errors.

There are limits, of course, to what can be shown by a single example – and neural network models of language acquisition are deeply controversial. But even with these caveats it should be clear how the tools provided by artificial neural networks provide a way of implementing the co-evolutionary approach to thinking about the mind. Using artificial neural networks to model cognitive tasks offers a way of putting assumptions about how the

mind works to the test – the assumption, for example, that the process of learning a language is a process of forming and evaluating hypotheses about linguistic rules. The test is, of course, in a sense rather contrived. As we saw earlier in the section, artificial neural networks are biologically plausible in only the most general sense. But, according to proponents of the artificial neural networks approach, to complain about this would be to misunderstand the point of the exercise. The aim of neural network modeling is not to provide a model that faithfully reflects every aspect of neural functioning, but rather to explore alternatives to dominant conceptions of how the mind works. If, for example, we can devise artificial neural networks that reproduce certain aspects of the typical trajectory of language learning without having encoded into them explicit representations of linguistic rules, then that at the very least suggests that we cannot automatically assume that language learning is a matter of forming and testing hypotheses about linguistic rules. We should look at artificial neural networks, not as attempts faithfully to reproduce the mechanics of cognition, but rather as tools for opening up novel ways of thinking about the mind and how it works.

We will be exploring the details and plausibility of this new way of thinking about the mind in subsequent chapters, but it is worth sketching out some of the broad outlines now to round off the presentation of the neurocomputational mind. An initial clue is provided by the central feature of the models of past tense acquisition that we have been considering. One of the key tenets of the neurocomputational approach to the mind is to downplay the role in cognition of explicit representations. Traditional approaches to the mechanics of cognition view cognition as a process of rule-governed manipulation of symbols. This is particularly clear on the representational picture, according to which all cognition involves transforming symbolic formulae in the language of thought according to rules operating only on the formal features of those formulae. This way of thinking about cognition rests, of course, on it being possible to distinguish within the system between the representations on which the rules are exercised and the rules themselves. But this distinction comes under pressure in artificial neural networks. The only rules that can be identified in these networks are the rules governing the spread of activation values forwards through the network and the propagation of error backwards through the network. There is nothing in either of the two models of past tense acquisition corresponding to the linguistic rule that the past tense is formed by adding the suffix '– ed' to the root of the verb. Nor are there any identifiable representations of the past tenses of irregular verbs. The network's "knowledge" of the relevant linguistic rules lies in the distribution of weights across all the connections in the entire network. There is no sense in which its "knowledge" that the past tense of 'go' is 'went' is encoded separately from its knowledge that the past tense of 'give' is 'gave'. There are no discrete representations within the system corresponding to the individual linguistic rules in terms of which we, as external observers, would characterize how the language works.

One crucial issue, then, is that although the network can plausibly be described as possessing various items of knowledge about how the past tense works for regular and irregular verbs, there are no discrete structures within the network corresponding to those items of knowledge. This marks a major point of difference between the neurocomputational picture of the mind and the pictures of the mind we have been considering up to now. Both the functional and the computational pictures take for granted that, for any personal-level propositional attitude, there must at some subpersonal level of explanation be a single discrete physical structure "standing in" for it. On the functional picture, this physical stand-in is whatever occupies the causal role defined by the personal-level propositional attitude. It is relatively easy to see how this might be applied to knowledge of the rule governing the past tense. That knowledge defines a certain causal role, with clearly definable inputs that are both psychological (the desire to refer to events in the past, for example) and linguistic (the roots of the relevant verbs) and clearly definable outputs (sentences featuring the appropriate past tense forms). The basic tenet of the functional picture is that, at some appropriate level of abstraction, there will be a physical structure occupying this causal role. The computational picture of the mind is committed, not just to there being a physical stand-in for the personal-level propositional attitude, but to that stand-in taking the form of a sentence in the language of thought – a sentence that is more or less synonymous with the characterization that a linguist might give of the rule in question.

Neither expectation *seems* to be met, however, in the artificial neural networks we have been considering. There do not appear to be any discrete, causally efficacious physical structures within the network that can be ascribed responsibility for the network's correct performance when it generates the appropriate past tense form of an input verb. Nor is anything sententially encoded within the network. It might be right to describe the network as knowing that the past tense of regular English verbs is formed by adding the suffix '– ed' to the root, but this is at best an imprecise characterization of the performance of the network as a whole – rather than a specification of something internal to the network and responsible for its performance. What is responsible for its performance is simply a complex pattern of weights across the network as a whole. The various items of knowledge that the network can be described as possessing are inextricably interlinked and *distributed* across the network as a whole.

Even though the example we have been considering is highly circumscribed, it may well incorporate a more general lesson. What we are talking about, after all, is how to understand a particular type of knowledge – knowledge of the past tense of English verbs. Knowledge is a propositional attitude, and what we have is a proposal for reconfiguring how we understand that propositional attitude. It is natural to wonder whether there might be room for a more extensive reconfiguration of how we think about propositional attitudes in general. Perhaps it is a mistake to think about

propositional attitudes as discrete physical structures. Perhaps propositional attitudes are realized in a distributed manner, more like the way in which information seems to be encoded in artificial neural networks. Perhaps we should view beliefs, for example, as dispositional properties of neural systems, rather than as discrete items within the neural economy. Much of the debate about the potential philosophical significance of artificial neural networks has focused on this question of distributed representation as part of a discussion of the architecture of cognition.

But there is also a wider issue, less frequently discussed. One thing that emerges quite clearly from the discussion of the neural network models of past tense acquisition is that our intuitive characterization of what the network knows is at best approximate. When we talk about what the network knows we are not really talking about a discrete state of the network as we would be on standard models of propositional attitudes. Those standard models maintain that propositional attitudes are discrete states with contents that we can readily identify. As soon, however, as one starts to question whether propositional attitudes are really best viewed as discrete states, the natural next step is to wonder about whether they have easily characterizable contents at all – and in particular to wonder about the extent to which we have the tools to provide an adequate characterization of those contents. The standard model of the propositional attitudes holds that we can characterize a propositional attitude by giving a sentence specifying its content – a sentence that says what it is that is believed or hoped or known. The assumption here, of course, is that a sentence can adequately capture this content, so that, for example, the content of a belief or a desire is given by a sentence that we would use to express it. But there are ways of thinking about how representation takes place in artificial neural networks that cast doubt upon this assumption. Suppose it is right that when we talk about the network knowing the rule that the past tense of regular English verbs is formed by adding the suffix '– ed' to the root of the verb, all we are really doing is describing its performance in a relatively coarse-grained way. This leaves us with an obvious question. How could the network's knowledge be characterized more accurately? What exactly is it that the network knows (given that we know what it actually *does*)?

In one obvious sense it may seem that the concept of knowledge is not really appropriate at all. The network does not really know anything, even once the training process is complete. Rather, it is in a complex dispositional state determined by the pattern of weights attaching to the connections between units. This complex dispositional state leads it to transform inputs into outputs in a certain way that we, from the outside and without much insight into the inner workings of the network, characterize in terms of knowledge of certain linguistic rules governing past tense formation. This suggestion may well not seem at all controversial. In fact, it may seem a very natural way of thinking about artificial neural networks. But recall that artificial neural networks are being deployed as a way of bringing some of the very general things we know about how the mind works to bear on how we

think about personal-level psychology and personal-level psychological explanation. Artificial neural networks are a way of implementing the co-evolutionary research methodology in pursuit of what I am terming the neurocomputational picture of the mind. And so, once again, one may be tempted to generalize. Perhaps the relation between our description of ourselves in terms of propositional attitudes and the reality of what is going on in our brains is much more like that between our characterization of a neural network in terms of propositional attitudes and the reality of what is going on inside that network than is imagined by the other pictures of the mind we have been considering.

Paul Churchland has made a radical suggestion here. When it comes to neural networks, we have to characterize their internal representational states in a way that captures the distributed nature of the processing involved. One way of doing this is to think of their states as locations in a multi-dimensional state space in which each dimension yields the range of possible activation values for each unit. So, for example, the state space of a network with 27 units contains 27 dimensions, and assigning to each unit a particular activation value uniquely determines a single point within that 27-dimensional space (a point that could equally be represented by a vector comprising an ordered sequence of those activation values). Once we have the notion of a state space clearly in view we can think of the sequence of states within a particular network as a trajectory within the state space.

Paul Churchland's radical suggestion is that we may end up characterizing our own representational states in similar terms. If and when we do, we will find out that the standard model of propositional attitudes is just as inaccurate when applied to us as when it is when applied to artificial neural networks. Here is the possibility he envisages:

> Suppose that research into the structure and activity of the brain, both fine-grained and global, finally does yield a new kinematics and correlative dynamics for what is now thought of as cognitive activity. The theory is uniform for all terrestrial brains, not just human brains, and it makes suitable conceptual contact with both evolutionary biology and non-equilibrium thermodynamics. It ascribes to us, at any given time, a set or configuration of complex states, which are specified within the theory as figurative "solids" within a four- or five-dimensional phase-space. The laws of the theory govern the interaction, motion and transformation of these "solid" states within that space, and also their relations to whatever sensory and motor transducers the system possesses.
>
> (Churchland 1981, p. 129)

How will this conceptual framework of solids within multi-dimensional phase space relate to the familiar framework of the propositional attitudes? Churchland's proposal is rather more measured than it is usually taken to be.

He suggests, in effect, that our vocabulary of propositional attitudes should be viewed as a simplification of the underlying multi-dimensional reality – a conceptual framework whose predictive and explanatory utility indicates not its accuracy, but rather the extent to which it abstracts away from and compresses the underlying complexity.

> According to the new theory, any declarative sentence to which a speaker would give confident assent is merely a one-dimensional *projection* – through the compound lens of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language – a one-dimensional projection of a four- or five-dimensional "solid" that is an element in his true kinematical state. Being projections of that inner reality, such sentences do carry significant information regarding it and are thus fit to function as elements in a communication system. On the other hand, being *sub*dimensional projections, they reflect but a narrow part of the reality projected. They are therefore unfit to represent the deeper reality in all its kinematically, dynamically and even normatively relevant respects.[13]
>
> (ibid.)

The picture here is striking, and nicely captures one way that the neurocomputational picture of the mind carries forward the co-evolutionary research approach. A method of modeling cognitive processes inspired by some general features of what we know about brain design generates a revision of some central features of how we think about ourselves and our cognitive processes.

Of course, it is a long leap from the highly circumscribed and artificial neural network models we have been considering to the radical reconfiguration of commonsense psychology proposed by Paul Churchland – and even those best disposed towards the neurocomputational picture of the mind might admit that we have nothing more than a striking and thought-provoking analogy. And indeed, proponents of the neurocomputational picture of the mind tend to emphasize that we know too little about how the brain works to engage in anything much more concrete than striking and thought-provoking analogies. Nonetheless, there is, I think, enough to go on to see how the neurocomputational picture of the mind might be developed. The following summarizes the main points.

13 Wernicke and Broca's areas are areas of the brain traditionally identified with linguistic capacities.

Checklist for the neurocomputational picture of the mind

- The neurocomputational picture of the mind rejects top-down models of vertical explanation in favor of a co-evolutionary model of how different levels of explanation interact.
- Proponents of the neurocomputational picture think that personal-level thinking about the brain needs to co-evolve with the sciences studying the neural dimension of cognition.
- Existing tools for studying the brain directly are not at the right level for studying how cognition takes place. The various types of neuro-imaging are too coarse-grained and single-neuron studies too fine-grained to explain how distributed patterns of activation across populations of neurons generate different types of cognitive activity.
- One way of avoiding this problem is through using mathematical modeling to generate artificial neural networks that obey some of the general principles of brain design and organization.
- Using artificial neural networks to model particular cognitive tasks (such as particular aspects of language acquisition) allow the co-evolutionary research strategy to be pursued.
- Artificial neural networks are particularly good at tasks involving pattern recognition.
- The only rules explicitly encoded into artificial neural networks are those governing how activation spreads through the network and the way in which the network deals with error.
- Representation in artificial neural networks is *distributed* across the units and the connections between them, rather than being encoded in discrete symbol structures.
- Thinking about the possibility that representations within brains might be distributed in a similar manner suggests ways of reconfiguring standard ways of thinking about propositional attitudes.
- Our practice of specifying propositional attitudes by giving a sentence specifying their content may turn out to be no more accurate than giving a sentence to characterize what is known by an artificial neural network.

6 Rationality, mental causation and commonsense psychology

- Real patterns without real causes
- How anomalous is the mental?
- The counterfactual approach
- Overview

Each of the four pictures of the mind discussed in previous chapters offers a different response to the *interface problem*. This is the problem of explaining how the commonsense psychological explanations that stand at the top of the hierarchy of explanation can be integrated with levels of explanation lower in the hierarchy. Each picture of the mind is driven by a different model of the vertical relations holding between personal and subpersonal levels of explanation. These different models are themselves intimately bound up with different ways of construing commonsense psychology. We looked at how the thesis of a radical discontinuity between personal and subpersonal explanation emerges from construing commonsense psychology as an essentially normative enterprise governed by distinctive standards of rationality and coherence. If, on the other hand, commonsense psychological explanation is understood to be governed primarily by causal laws, rather than by normative principles of rationality, then a functional understanding of the interface between personal and subpersonal levels becomes attractive. A particular way of thinking about what it is required for propositional attitudes to be causally explanatory in virtue of their content has inclined many theorists towards some version of the representational picture of the mind and the language of thought hypothesis. And we saw how the thought that personal-level psychological explanation of behavior cannot be understood independently of our subpersonal understanding of how the mind works leads to different versions of the neurocomputational picture of the mind.

Most philosophers of psychology hold commonsense psychological explanations to be causal explanations on a par with causal explanations at the various subpersonal levels in the hierarchy of explanation (and, of course, with those in completely different domains of the social and natural sciences). This standard picture of psychological explanation has two key elements. The first is that commonsense psychological explanations can only be causal if there are causally efficacious internal items corresponding to (realizing, or serving as the vehicles of) the propositional attitudes cited in those explanations. The second is that the causal dimension of commonsense

psychological explanations requires the existence of causal laws governing the inter-relations between psychological states and between psychological states and behavior. The standard picture of psychological explanation that emerges from these two assumptions has been challenged by an influential minority of theorists. These theorists defend different versions of the autonomy picture.

Three different and independent challenges to the standard picture of psychological explanation will be explored in this chapter. The first comes from Daniel Dennett, who has developed a distinctive understanding of commonsense psychology. Commonsense psychological explanations track genuinely existing patterns in the behavior of organisms (and cognitive systems more generally), but not in a way that requires the existence of independently identifiable and causally interacting physical structures. According to Dennett, commonsense psychological explanations can be true without being causal *in anything like the standard sense assumed by philosophers* – and, in particular, without there being identifiable and discrete inner items corresponding to individual propositional attitudes. Dennett's position and the motivations for it will be the subject of section 6.1. In section 6.2 we consider another influential line of argument in support of the autonomy picture. The question of whether commonsense psychological explanations are causal is connected to (although not equivalent to) the question of whether the generalizations of commonsense psychology are strict, law-like generalizations. Autonomy theorists such as Davidson have argued that the generalizations of commonsense psychology are essentially normative in a way that precludes them from being strict causal laws (and hence that there is a fundamental error in functional and representational attempts to characterize the mental in causal terms). This conception of commonsense psychological generalizations is at the heart of the autonomy theory's insistence on the incommensurability of personal and subpersonal levels of explanation. The final section of the chapter (6.3) considers a challenge to the basic assumption that causal explanation requires causal laws. According to proponents of the counterfactual approach to mental causation and psychological explanation, there can be mental causation without causal laws because all that is required for a causal relation to hold is the truth of certain counterfactual statements about what would have happened in relevantly different circumstances. The counterfactual approach provides a further way of developing the picture of personal-level explanation as fundamentally independent of subpersonal-level explanation.

6.1 Real patterns without real causes

Daniel Dennett's views on the nature and aims of psychological explanation have evolved in the thirty or so years since *Content and Consciousness* (Dennett 1969). But Dennett has consistently resisted what has become a standard inference among philosophers of mind and philosophers of psychology. This

is the inference from the usefulness and accuracy of psychological explanation to the existence of *causally efficacious* internal items corresponding to the beliefs and desires cited in those explanations. What makes a psychological explanation useful and accurate, so the argument goes, is its truth, and that truth consists in correctly identifying the beliefs and desires responsible for generating the behavior in question. Responsibility here is to be understood in causal terms. So we are led to the demand for causally efficacious internal items, generally understood to be neurophysiological states of one kind or another. Much of contemporary philosophy of mind is occupied with the question of how exactly these internal items relate to the personal-level states cited in psychological explanations. Are they identical? If so, between what does the identity hold? Is it an identity holding between individual neurophysiological states and individual psychological states? Or is it an identity holding across types of state?[1] Perhaps, instead, the relation should be understood as one of realization, rather than identity, in the way that functionalism suggests?[2]

Dennett's initial resistance to this line of argument focused on the first stage – on the claim that we need to understand the predictive utility of psychological explanations in terms of their truth. At various points in his earlier writings (e.g. Dennett 1981) he developed a position that has struck many as a form of instrumentalism about psychological explanations. This instrumentalism effectively turns the standard argument on its head, maintaining that there is nothing more to the truth of psychological explanations than their predictive utility. Dennett once suggested that all there is to being a believer is behaving in ways that are usefully explicable according to what he called the intentional stance – that is to say, within the personal-level framework of commonsense psychology. Behavior is usefully explicable according to the intentional stance when bringing to bear the machinery of belief–desire explanation permits successful predictions that are not available when one considers the system in question from the physical stance or from the design stance (see section 2.1). So, for example, it is useful to consider a chess-playing computer as having a desire to win and certain beliefs about tactics and strategy. This gives us purchase on ways it might behave that would otherwise be unavailable. In contrast, however, there is little mileage to be gained from treating a thermometer as an intentional system. We can fully understand what a thermometer is going to do once we understand the general principles governing its design. No predictive power is added by taking it to have a desire to track the ambient temperature.

From the instrumentalist point of view, the standard argument falls at the first hurdle, since explanation according to the intentional stance is not accountable to independent standards of truth or falsity. Its very applicability secures its truth. Once we have established that applying the intentional

1 For the distinction between token-identity and type-identity see section 3.2.

2 Different versions of the functional picture are outlined in sections 3.3 and 3.4.

stance in a given situation has a predictive or explanatory pay-off, there is no further question to be asked about whether the explanation or prediction is true – and so no question to be asked about whether or not the system in question really does have the beliefs and desires cited in the explanation or prediction. Few readers of Dennett have been satisfied by this instrumentalism, however, and more recently Dennett himself has moved away from instrumentalism towards what he calls a "mild realism" about propositional attitudes (Dennett 1991a, p. 30). The difference between instrumentalism and mild realism is, in essence, a view about the *truth-aptness* of commonsense psychological explanations. Mild realism permits a more robust sense in which psychological explanations can be evaluated for truth or falsity. Dennett thinks that the truth of belief–desire explanations consists in their tracking genuinely existing patterns in the behavior of the organisms and systems to which those explanations are applied. These patterns, which Dennett calls "real patterns", are not observer-dependent (in the way that ascriptions of propositional attitudes seemed to be on Dennett's earlier instrumentalism). There is a genuine fact of the matter as to whether they hold or not. These independently existing real patterns provide the truth-makers for psychological explanations – they are the beliefs and desires cited in the explanation. And the existence of those patterns is, of course, what makes commonsense psychological explanation and prediction effective.

Dennett's mild realism gives us a way of accepting the first two stages of the standard line of argument but rejecting the third. It explains the predictive utility of commonsense psychological predictions in terms of their truth and it offers a genuine way of understanding what that truth consists in, namely, in the existence of beliefs and desires understood as patterns in the behavior of cognitive systems. Yet there is no need to appeal to causally efficacious inner items in order to explain the success of commonsense psychological explanations. Those explanations are successful just when they latch on to real patterns.[3]

Dennett's real patterns hold over *emergent* properties of intentional agents and cognitive systems. They are patterns in the behavior of the agent or system as a whole and cannot be reduced to, or understood in terms of, the operation of parts of the agent or system. This is part of what makes Dennett an autonomy theorist and allows him to deny that personal-level facts about the behavior of the system as a whole can be understood in terms of facts about inner states or modules at the subpersonal level. The notion of

3 Dennett is not denying that commonsense psychological explanations are causal. He just thinks that causal explanations do not require causally efficacious inner items: "If one finds a predictive pattern of the sort just described one has ipso facto discovered a causal power – a difference in the world that makes a subsequent difference testable by the standard empirical methods of variable manipulation" (1991a, p. 43, n.21). As this passage makes clear, Dennett has a slimmed-down notion of causation, similar in spirit to the counterfactual approach introduced in Chapter 4 and discussed further in section 6.3.

an emergent pattern is an important one. Dennett illustrates it with a beautifully simple example developed by the mathematician John Conway – Conway's Game of Life.

The Game of Life is an example of a *cellular automaton*, which is a mathematically defined model of an artificial universe defined over an array of cells and governed by a simple set of "physical" laws. The array of cells defines the space (which can be of any number of dimensions) of the artificial universe. Each of the cells in the array has a small number of possible states. In the simplest cellular automata a cell can be on or off. The evolution of a cell from one state to another is governed by a *transition function* that determines the state of any given cell as a function of the states of its neighbors. The process of evolution operates over discrete time-intervals. The transition function is purely local in the sense that each cell has information only about its immediate neighbors. The aim of cellular automata is to investigate the complex behavior that can emerge from the very simple rules specified by the transition function.

In the Game of Life the array is a two-dimensional grid, rather like a large chessboard (a typical grid size might be 1000×1000). Each cell is connected to the eight cells with which it is in contact (the four cells with which it shares an edge and the four cells touching at the corners). Each cell can be either On (have value 1) or Off (have value 0). The transition function governing the state of the cells is very straightforward. It is composed of three rules:

1 If a cell is off at time t and three of its neighbors are on, then at time $t + 1$ the cell is switched on.
2 If a cell is on at time t and either two or three of its neighbors are also on, then at time $t + 1$ the cell will remain switched on.
3 If any other configuration holds at time t then the cell will be off at time $t + 1$.

One can think of a cell as requiring a certain number of neighbors to come into existence, and as dying when its environment becomes either overpopulated (with more than three neighbors) or underpopulated (with fewer than two neighbors). These three rules cover all the possible situations and, given their apparent simplicity, one might have thought that the behavior of the system would be very predictable and in fact rather tedious.

As it turns out, however, the Game of Life is anything but predictable. The behavior of the system depends upon the initial configuration of cells and slightly different starting configurations will yield drastically different outcomes. There are no techniques that allow us to predict what the result will be for any initial configuration – other than actually running a simulation to see what happens.[4] And this is the case even though the system is

4 There are many simulations of the Game of Life (and other cellular automata) available on the Internet. A useful collection of links to downloadable simulations will be found at http://radicaleye.com/lifepage.

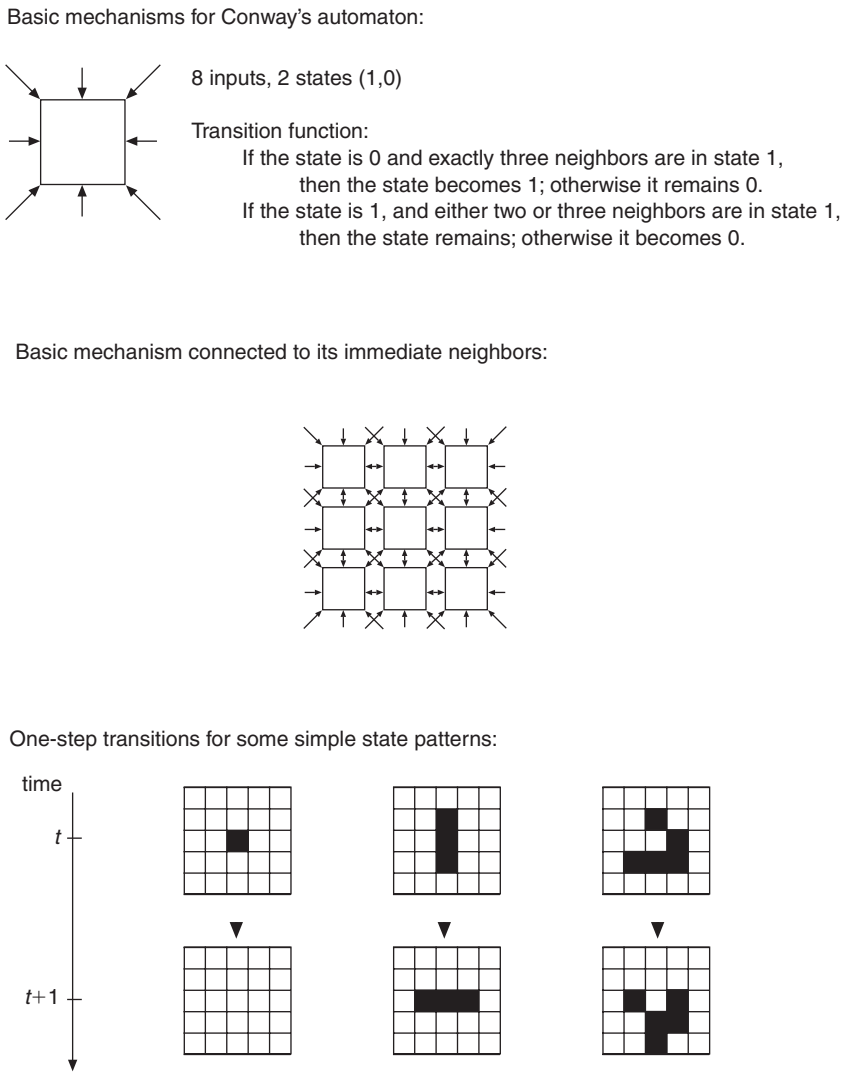

Figure 6.1 Conway's automaton (source: Holland (1998, p. 137)).

totally deterministic, in that each configuration of the system at a given time is fixed by the transition function and the state of the system at the preceding time. Figure 6.1 illustrates the design of the Life World and gives a sense of how some basic configurations evolve over a single time-interval.

What is interesting about the Life World is that a certain number of basic patterns reappear in the evolution of a very great number of initial configurations. The most basic and best known of these is the so-called

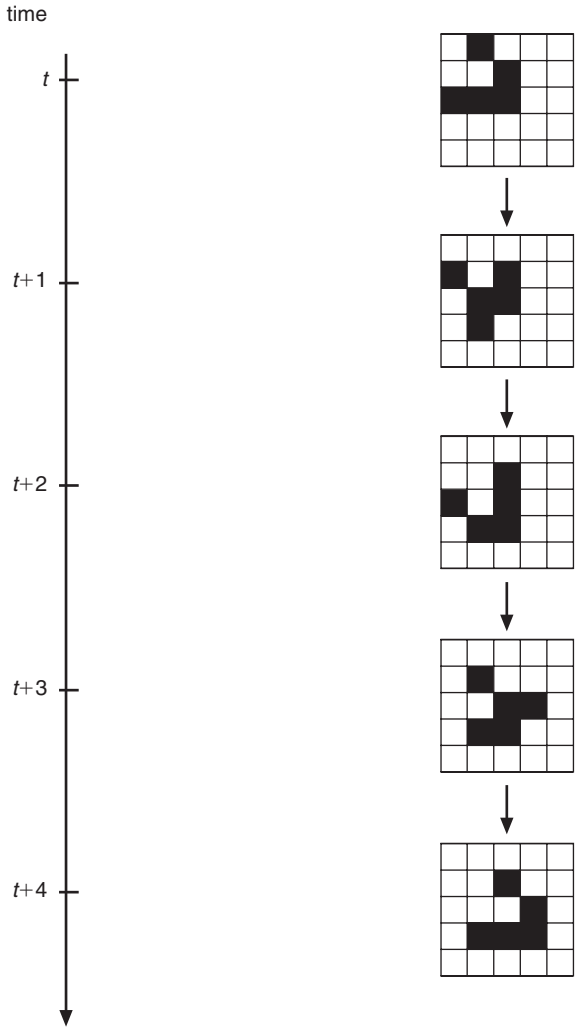

Figure 6.2 Successive transitions of the glider state pattern in Conway's automaton (source: Holland (1998, p. 138)).

glider pattern, a configuration of five cells that changes shape in a regular way. As Figure 6.2 illustrates, the glider pattern reproduces itself in four time-intervals, reappearing one square diagonally down and to the right of its starting position. Provided that nothing interferes with it, the glider pattern will (as its name suggests) glide diagonally rightwards down the array.

The glider pattern is just one of the patterns that emerge in the Life World. An hour or two experimenting with a computer simulation of the

Life World and the many databases of initial configurations available on the Internet will turn up others. The glider pattern is sufficient, however, to make the basic point about emergence. As Dennett points out, we can shift levels of description by talking about gliders moving across the array to yield a significant economy in the amount of detail required to characterize the evolution of the Life World. If we introduce gliders into our Life World ontology then we save ourselves the trouble of describing what is going on in terms of a cell-by-cell description. All we need do is identify the glider's starting and end positions. But by doing this we are introducing new patterns into the system. At the basic level of description (the level of description at which the transition function is given), there is no movement at all in the Life World. The Life World is composed solely of cells switching state from On to Off. By switching to talk of gliders, however, we introduce movement into the system. This movement is, in Dennett's terms, an emergent property. It tracks a real pattern (the recurring pattern in cells switching state that we describe as a glider), but the pattern that it tracks does not actually display the higher-level property. The movement identified at the higher level tracks a genuinely existing pattern at the lower level, but the pattern at the lower level is not a pattern of movement. At the level of individual cells there is no movement – only individual cells switching off and on.

The Game of Life illustrates that there can be patterns that are perfectly real, even though they are not what they appear to be – that is to say, even though their existence is determined by lower-level processes that lack certain fundamental properties of the higher-level pattern. It is, of course, a long way from the Life World to the behavior of intentional systems, but we can already see how the first stage in Dennett's strategy is supposed to work. Dennett thinks that there are real patterns in the personal-level behavior of intentional systems that correspond to what is going on at the subpersonal level in the way that glider patterns in the Life World correspond to what is going on at the cell-by-cell level of description. If he can secure this parallel, then it should secure him against the charge of instrumentalism. The patterns in the behavior of gliders are not purely observer-dependent. They are robust, genuinely existing properties of the Life World – just not ones discernible at the lowest level of description. The patterns have truth-makers. There is a particular sequence of configurations of cell-states that makes true the claim that a glider is moving across the grid – and indefinitely many other configurations of cell-states that will falsify the claim.

But how are we to move from the Life World to the far more noisy and complex world of intentional agents? The basic point about the Life World is that there is no straightforward mapping from higher-level descriptions to lower-level descriptions. The higher-level description can be true even though there is no possibility of identifying the items over which it is defined at the lower level of description. Dennett draws a direct analogy between the ontology of the Life World and the ontology of propositional

attitude psychology in this respect. It makes no more sense to look for sub-personal-level items corresponding to beliefs and desires than it makes sense to look for moving gliders in a cell-by-cell description of the Life World. The point emerges very clearly in the following comments Dennett makes about Fodor:

> For Fodor, an industrial-strength Realist, beliefs and their kin would not be real unless the pattern dimly discernible from the perspective of folk psychology could also be discerned (more clearly, with less noise) as a pattern of structures in the brain. The pattern would have to be discernible from the different perspective of a properly tuned *syntactoscope* aimed at the purely formal (non-semantic) properties of Mentalese terms written in the brain. For Fodor, the pattern seen through the noise by everyday folk psychology would tell us nothing about reality, unless it, and the noise, had the following sort of explanation: what we discern from the perspective of folk psychology is the net effect of two processes: an ulterior, hidden process wherein the pattern exists quite pure, overlaid and partially obscured by various intervening sources of noise: performance errors, observation errors, and other more or less random obstructions.
>
> (Dennett 1991a, pp. 42–43)

Were Fodor to be contemplating the Life World, he would, Dennett thinks, deny that it really contains moving gliders. From the perspective of "industrial-strength realism" there has to be an isomorphic lower-level pattern for any genuinely existing higher-level pattern. Dennett is trying to show, in contrast, that the genuinely existing patterns of commonsense psychology can be properly grounded (can have truth-makers) without any such isomorphism.

But how can there be grounding for the patterns of commonsense psychology without isomorphism? Dennett's response to this question is rather elliptical. Here is what he says:

> But how *could* the order be there, so visible amidst the noise, if it were not the direct outline of a concrete orderly process in the background? Well, it *could* be there thanks to the statistical effect of very many concrete minutiae producing, as if by a hidden hand, an approximation of the "ideal" order.
>
> (ibid.)

There are two ideas here that need to be separated out. The first is the idea that a multitude of lower-level causes might contribute to producing a higher-level pattern. This is precisely what the Life World illustrates so clearly. But in applying this general idea to commonsense psychology Dennett introduces a further idea. This is the idea that the patterns of com-

monsense psychology are best viewed as "approximations to an ideal order". To understand this, we need to think back to how psychological explanation is characterized by autonomy theorists such as McDowell and Davidson. These theorists see psychological explanation as governed by norms of rationality that determine, not how people as a matter of fact do behave, but rather how they ought to behave. Psychological explanations work (on this view) by fitting observed behavior into a framework in which it makes sense from the agent's point of view (given what that agent wants to achieve and his information about the world). One way of putting this would be to say that we interpret people's behavior by showing how it approximates to an ideal of rationality – the generalizations that we use in explaining their behavior are norms to which we see them as aspiring, rather than laws under which their behavior is to be subsumed.

If we view psychological explanation in this way, then it makes sense to ask why it is that people behave in ways that give the impression they are aspiring to norms of rationality. One common answer to this question, emerging very clearly in Fodor's writings on intentional explanation, is that this appearance of rationality is the result of subpersonal states interacting in ways that correspond to the basic principles of rationality (in virtue of the interdependence of syntax and semantics that we examined in Chapter 4). Dennett, however, offers a fundamentally different explanation. The appearance of rationality emerges as "the statistical effect of very many concrete minutiae producing, as if by a hidden hand, an approximation of the 'ideal' order".

It is no use looking to the Life World to help understand how this "invisible hand" is supposed to work. The patterns discernible in the Life World are not approximations to an ideal order. The movement of a glider is no more and no less than the movement of a glider. The analogy that Dennett himself offers is how the wings of birds work according to the principles of aerodynamics without those principles being explicitly represented in them. But, while this example does illustrate the difference between acting in accordance with a principle and explicitly following a principle, it is too far removed from the sphere of psychological explanation to be much help. A better example of what Dennett is getting at comes, I think, from the approach to explaining animal behavior known as *optimal foraging theory* (for an introduction, see M. S. Dawkins 1995, Chapter 2, and Krebs and Kacelnik 1991, and Parker and Maynard Smith 1990, for more advanced surveys).

It is possible to model certain aspects of animal behavior by making the heuristic assumption that animals are performing complex cost–benefit calculations. Here is how Krebs and Kacelnik (1991) describe the bare bones of the framework they propose for studying patterns of animal behavior, such as those displayed by a robin seeking food (foraging):

> We shall use the metaphor of the animal as a 'decision-maker'. Without implying any conscious choice, the robin can be thought of as 'deciding'

> whether to sing or to feed, whether to feed on worms or on insects, whether to search for food on the grass or on the flower bed. We shall see how these decisions can be analyzed in terms of the costs and benefits of alternative courses of action. Costs and benefits are ultimately measured in terms of Darwinian fitness (survival and reproduction), and may, in many instances, be measured in terms of some more immediate metric such as energy expenditure, food intake or amount of body reserves. Analyzing decisions in terms of their costs and benefits cannot be done without also taking into consideration physiological and psychological features that might act as constraints on an animal's performance. The fitness consequences of decisions, and the various constraints that limit an animal's options, can be brought together in a single framework using optimality modeling.

The guiding assumption of optimal foraging theory is that animals should optimize the net amount of energy obtained in a given period of time. Acquired energy is the benefit in the cost–benefit analysis. In the case of a foraging bird, for example, faced with the "decision" of whether to keep on foraging in the location it is in or to move to another location, the costs are the depletions of energy incurred through flight from one location to another and during foraging activity in a particular location. The cost–benefit analysis can be carried out once certain basic variables are known, such as the rate of gaining energy in one location, the energy cost of flying from one location to another and the expected energy gain in the new location. It turns out that optimality modeling makes robust predictions of foraging behavior in birds such as starlings (*Sturnus vulgaris*) and great tits (*Parus major*).

Of course, as Krebs and Kacelnik make plain in the quoted passage, there is no suggestion that the great tits or starlings really are carrying out complex calculations about how net energy gain can be maximized within a particular set of parameters and background constraints. It is a crucial tenet of optimal foraging theory that the optimizing behavior is achieved by the animal following a set of relatively simple rules of thumb or heuristics, which are most probably innate rather than learned. So, for example, a great tit might be hard-wired to move on to the next tree after a certain number of seconds spent unsuccessfully foraging in one tree. Evolution has worked in such a way (at least according to the proponents of optimal foraging theory) that foraging species have evolved sets of heuristic strategies that result in optimal adaptation to their ecological niches. This optimal adaptation can be mathematically modeled, but the behaviors in which it manifests itself do not result from the application of such a theory – any more than, to return to Dennett's own example, a bird's ability to fly reflects any mastery on its part of the basic principles of aerodynamics.

The possibility is opening up of interpreting human behavior and practical decision-making as driven by heuristics and rules of thumb that

"approximate to an ideal of rationality" in much the same way that the simple heuristics driving foraging behavior approximate to the ideal of rationality determined by a cost–benefit analysis. What sort of heuristics and rules of thumb might these be? We can get some clues from two related sources – empirical studies in the psychology of reasoning and proposals that have been made about the evolution of cognition. Both of these can be understood against the background of some of the anomalies that researchers have found in subjects' grasp of some formal principles of reasoning.

Researchers in the psychology of reasoning have produced robust evidence that subjects frequently reason in ways that contravene some basic principles of deductive logic and probability theory. Here are some examples:

- A study carried out in 1977 (Rips 1983) showed that the only basic conditional argument that their subjects could apply reliably was *modus ponens*. There was a noticeable tendency to affirm the consequent and deny the antecedent. Twenty-one percent of the subjects said that an argument that denied the antecedent would always be valid – while the figure was 23 percent with affirming the consequent. Also striking is the fact that 43 percent failed to see that *modus tollens* arguments were always valid.[5]

  Another good example of failure in elementary deductive reasoning is to be found in the selection task experiments carried out by Wason and Johnson-Laird (1972). The subjects were presented with four cards (Figure 6.3) and then asked to evaluate the conditional 'if there's a circle on the left then there's a circle on the right' by saying which cards they would have to see completely in order to answer the question. The answer is that cards (a) and (d) must be unmasked. Unfortunately, only five out of 128 college students realized this. Almost all the 123 who got it wrong failed to see the need to turn (d) over. This is failing to see the equivalence of a conditional, 'if p then q', with its contrapositive, 'if $\sim q$ then $\sim p$'.
- It is a basic principle of statistics that the probability of a given sample being representative of the population from which it is drawn varies in proportion to the size of the sample. A large sample is less likely than a small sample to diverge from the mean for the population as a whole. This is the so-called law of large numbers. Yet people tend to judge even small samples to be highly representative (Tversky and Kahneman 1971).

5 To refresh the memory, a *modus ponens* inference derives 'q' from the premises 'if p then q' and 'p', while a *modus tollens* inference derives '$\sim p$' from 'if p then q' and '$\sim q$'. Both of these are valid inferences – they will never lead from true premises to a false conclusion. Each, however, has a counterpart that is superficially similar but invalid. The fallacious counterpart of *modus ponens* is the fallacy of affirming the consequent – deriving 'p' from 'if p then q' and 'q'. The fallacious counterpart of *modus tollens* is the fallacy of denying the antecedent – concluding '$\sim q$' from 'if p then q ' and '~p'.

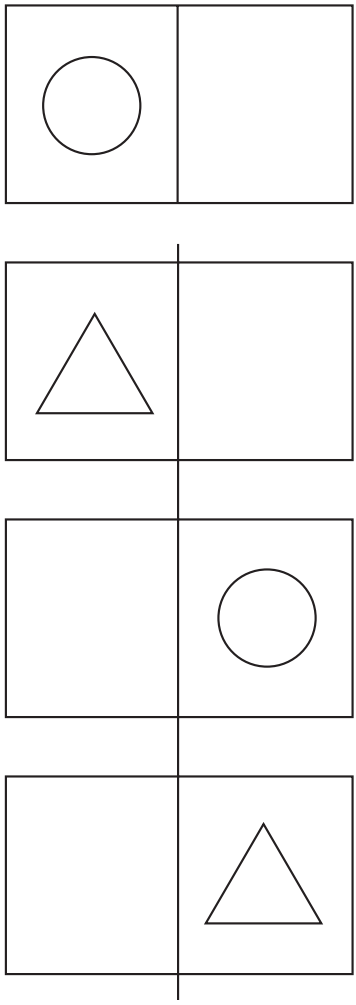

Figure 6.3 Figure from Wason selection task.

- A fundamental principle of the probability calculus is that the probability of a conjunction cannot be greater than the probability of one of its conjuncts.[6] But subjects regularly commit the conjunction fallacy of assigning a higher probability to a conjunction than to one of its conjuncts (Tversky and Kahneman 1983).

6 This follows from the so-called extension rule of probability – that if the extension of A includes the extension of B then $P(A) \geq P(B)$.

There has been considerable debate within both philosophy and psychology about whether these experiments (and the many others like them) illustrate that human beings are in some sense deeply irrational.[7] Putting this debate to one side, one of the most fruitful aspects of this research has been the range of accounts it has generated of how people actually go about reasoning. The idea of heuristics or rules of thumb has been at the forefront of many of these accounts (Tversky and Kahneman 1974; Gigerenzer *et al.* 1999). Consider, for example, the problems that people have with the law of large numbers – problems that manifest themselves in people judging (contrary to basic principles of statistics) that small samples are just as indicative of what one will find in a population as a whole as large samples. These difficulties are perfectly understandable if subjects are in fact employing what has come to be known as the *representativeness heuristic*, in which the probability of a sample being typical of a population is taken to be a function of the representativeness of the sample. A small sample can be just as representative of a population as a large sample. Another heuristic that has been proposed as central to practical reasoning is the *availability heuristic*, according to which the probability of an event is judged to be a function of the ease with which examples can be brought to mind. One can see, for example, how the availability heuristic might lead someone to think that a conjunction is more likely than one of its conjuncts. Suppose, for example, that one is asked whether Mr Jones is more likely to be a farmer or a farmer with a four-wheel drive vehicle. The probability is obviously higher that he is a farmer, since there are many farmers who do not have four-wheel drive vehicles. But it may well be easier to bring to mind an example of a farmer with a four-wheel drive than one without a four-wheel drive. Hence the ease with which people fall into the so-called "conjunction fallacy" of thinking that a conjunction can be more probable than one of its conjuncts.

The availability and representativeness heuristics are examples of so-called "fast-and-frugal" heuristics – short-cuts in problem solving and decision-making that save computational time while getting things right enough of the time to be adaptive (Gigerenzer *et al.* 1999). The research program into reasoning heuristics has been given an impetus by a set of influential proposals in evolutionary psychology (Cosmides and Tooby 1992). There is wide support among evolutionary psychologists for the hypothesis that the evolution of cognition involved the emergence of a set of highly specialized cognitive modules, each dedicated to solving a particular adaptive problem. These cognitive modules (which should not be assimilated to Fodorean modules of the type discussed in earlier chapters) do not share general principles of reasoning.[8] On the contrary, they each work with a set of specialized rules that evolved explicitly to solve the adaptive

7 See, for example, Stein (1996) and Evans and Over (1996) for overviews of the philosophical and psychological debates respectively.

8 Darwinian modules are discussed further in section 8.4.

problems faced by our hominid ancestors. Imagine a social group of hominids. This group will work best if it does not contain any free-riders (a free-rider is someone who reaps the benefits of communal existence without making their own contribution) and so it becomes a pressing matter to identify the free-riders. In response to this pressing evolutionary need, it has been proposed, early hominids developed a particular facility for a certain type of conditional reasoning that allows the detection of cheaters and other types of free-riders. This is reasoning involving conditionals of the form "if *p* then *q*" where the conditional fails to hold just when someone takes a benefit without paying an appropriate cost. What is interesting about this Darwinian algorithm (as it has been dubbed) is that it allows us to make sense of some curious experimental phenomena that emerged in the extensive research done on the Wason selection task (see above). Studies have shown that when the Wason selection task is reformulated as a task that requires detecting a cheater or a free-rider, subjects perform much better on it (Cosmides 1989). The precise interpretation of these data is a matter of some controversy (see Evans and Over 1996, Chapter 4, and section 8.4 below for more discussion), but they at least open up the possibility that a significant part of everyday reasoning may involve content-sensitive rules (rather than the abstract, formal rules of deductive logic).

Suppose, then, that human beings do in fact employ a range of heuristics, rules of thumb and content-sensitive reasoning principles of reasoning. In everyday life (as opposed to the artificial protocols of the psychologist of reasoning), these short-cuts work very well. Well enough, in fact, to give the appearance that reasoners and agents are in fact conforming to the norms of rationality in the way that commonsense psychological explanation demands (at least according to Dennett, McDowell, Davidson and other like-minded theorists). Would this not then meet Dennett's description of "very many concrete minutiae producing, as if by a hidden hand, an approximation of the 'ideal' order"?

The operation of reasoning heuristics, rules of thumb and specialized Darwinian algorithms can give the appearance that agents are following the normative principles of rationality, even though they are at best merely approximating to those principles and in no sense explicitly following them. In fact, as the experimental studies (arguably) show, when those same agents do set out explicitly to follow the very normative principles that we make use of in predicting their behavior, they are not very good at it. So, even though there are certain patterns of rationality discoverable in their behavior (patterns that in some sense approximate to an ideal rationality), the corresponding rules and principles are neither built into the system not explicitly followed by it. This is the key to understanding why Dennett thinks that commonsense psychological explanations can have truth-makers (can be properly grounded) without there being any causally efficacious inner items at the subpersonal level corresponding to the beliefs and desires cited in those explanations.

Let us look in more detail at how the process works. A commonsense psy-

chological explanation or prediction attributes to an agent a set of beliefs and desires that jointly make a given action comprehensible (either retrospectively, when we are dealing with explanation, or in anticipation, when we are making a prediction). In attributing a set of beliefs and desires to an agent, one is thereby identifying corresponding patterns in their behavior – because one is incurring commitments as to how they will behave in relevantly similar circumstances, in different circumstances, and so on. The explanation or prediction will be true just if the agent's behavior really does display those patterns – if the commitments incurred are in fact borne out. This is a personal-level fact that has a subpersonal truth-maker. What grounds the patterns displayed at the personal level is a complex of mechanisms, rules and algorithms at the subpersonal level. So, we have truth-makers without isomorphism.

Dennett's idea of real patterns that are non-isomorphically grounded is a useful corrective to some very deeply engrained ways of thinking about the relation between personal and subpersonal levels of explanation. He is surely correct to emphasize that we should not always be looking for structural isomorphism between personal and subpersonal levels of explanation and that the patterns we detect at the personal level may be produced by all sorts of subpersonal mechanisms acting "blindly". The foraging bird is in many instances a better model for thinking about how our explanations work than the billiard balls that so often dominate philosophers' thinking about causation. But one might wonder just how universally applicable Dennett's model can be. Is the standard model of beliefs, desires and other propositional attitudes as causally efficacious inner items really a *complete* myth? Could it really be the case that *all* instances of apparently rational behavior are really approximations to an ideal rationality blindly generated by subpersonal mechanisms and heuristics?

Dennett needs to explain what entitles us to describe a particular pattern as genuinely existing, where a genuinely existing pattern is one that would hold independently of its being identified by any observer. A genuinely existing pattern must be detected, rather than created, by explanations that appeal to it. Dennett is sensitive to this requirement and offers an account of what makes a pattern a real pattern. Suppose we have a description of the phenomenon that we are setting out to explain – which may be, for example, a particular person's action. That description is our *explanandum* – what it is that we are trying to explain. Our explanation works by giving another description. This description picks out a pattern under which the *explanandum* falls – in the case we are considering, a pattern of behavior that subsumes the particular action we are trying to explain (the *explanans*). So, we effectively have two descriptions of the same behavior. One description picks it out in neutral terms (or relatively neutral terms) while the other characterizes it as falling under a particular pattern. A particular relation has to hold between those two descriptions for there to be a real pattern in play. According to Dennett, we have a real pattern whenever the second description is more

efficient than the first. The efficiency of a description is a function of the amount of detail it involves – the less detail, the more efficient. When we see events as instantiating patterns, we make a trade-off. We lose some of our original information about the event, but by doing so we make it possible to see what it might have in common with other events. Identifying patterns is a process of abstraction. Dennett's suggestion is that a real pattern can be identified whenever such a process of abstraction is possible.

We can illustrate the point in terms of the image on a computer screen. The image can be described by a process that gives a value to every pixel – this highly detailed description corresponds to our *explanandum*. This pixel-by-pixel description is a bit-map (and can be compared to a cell-by-cell description of a particular stage in the Game of Life). The efficiency of pattern can be understood in terms of the number of bits it employs. There is a real pattern in the image on the screen when there exists a description that requires fewer bits than the bit-map. As Dennett puts it, "a pattern exists in some data – is real – if *there is* a description of that data that is more efficient than the bit map, whether or not anyone can concoct it" (1991a, p. 34). These patterns are, he stresses, observer-independent. Descriptions can be true of something whether or not anyone formulates them.

The analogy with commonsense psychology is clear. Commonsense psychology offers extremely efficient tools for characterizing behavior. It achieves this efficiency by sacrificing detail. Commonsense psychology abstracts away from the subpersonal "bit-map", gaining usefulness at the cost of losing information. The possibility of such abstraction, Dennett maintains, explains the existence of real patterns in people's behavior.

A potential problem, however, is that this conception of what makes a pattern real seems too generous. It appears to yield too many genuinely existing real patterns, given that any piece of behavior can be abstractly described in numerous different ways. The problem is not just that there are different degrees of abstraction – and hence that one can have abstract descriptions with more or less "noise" in them. Dennett is quite right to stress that this is not a problem. We require different levels of detail for different purposes. Some explanations will require a relatively low-level description ("he ate the cheese because he was hungry") that abstracts away from most of the details of the case. In other contexts one needs an explanation that is much finer-grained ("he ate the cheese because he had wanted for a long time to experiment with unpasteurized Camembert"). The degree of abstraction required is determined by pragmatic considerations. Some of these are to do with what are taken to be relevant alternatives. If the issue is why the person ate the cheese rather than going for a walk, then the mention of hunger will suffice. But if the issue is why the person ate the cheese rather than any of the other delicacies available, then more detail is required. There are other trade-offs to be made between predictive accuracy and convenience. It can be advantageous to be "quick and dirty", allowing the possibility of error in order to avoid the diminishing returns that come

when one imposes too high a standard of accuracy. It seems, then, that there is no single appropriate level of abstraction. The necessary degree of abstraction is fixed by the requirements of the situation and the explanatory context.

Characterizing a single event in different ways at different levels of explanation does not give us conflicting explanations. Somebody might be both hungry and keen to try unpasteurized Camembert. In some contexts it will be appropriate to mention one factor rather than another. The real problem with Dennett's criterion for what makes something a genuine pattern, however, is that it leaves open the possibility of finding patterns that really do conflict. This is something that Dennett is quite prepared to admit:

> I see that there could be two different systems of belief attribution to an individual which differed *substantially* in what they attributed – even in yielding substantially different predictions of the individual's future behavior – and yet where no deeper fact of the matter could establish that one was a description of the individual's *real* beliefs and the other not. In other words, there could be two different, but equally real, patterns discernible in the noisy world. The rival theorists would not even agree on which parts of the world were pattern and which noise, and yet nothing deeper would settle the issue. The choice of a pattern would indeed be up to the observer, a matter to be decided on idiosyncratic pragmatic grounds.
>
> (ibid., p. 49)

Dennett is not reneging on his earlier insistence that real patterns are observer-independent. The observer has to decide, not what patterns there are, but rather which of the independently existing patterns he wants to emphasize.

Dennett is happy to embrace the conclusion that there is no fact of the matter about whether one explanation is better than another. Even when the candidate patterns are actually in conflict (as opposed to simply involving different degrees of abstraction), there is no fact of the matter about which pattern we should use in explaining a given action. There is no sense in which an explanation that invokes one set of beliefs and desires can be more or less well grounded than one that invokes a different and incompatible set of beliefs. Each exploits a genuinely existing pattern. Many theorists, however, will be unwilling to accept this conclusion, not least because it appears to have the consequence that an agent can simultaneously have radically inconsistent beliefs. As we have seen, if Dennett is to avoid instrumentalism, he needs to offer a robust sense in which attributions of beliefs and desires can be true. This is what his account of real patterns is intended to achieve. Combining this with the thesis that there is no fact of the matter about which of a number of conflicting patterns is the real pattern, however, yields the conclusion that there are beliefs and desires corresponding to every genuinely existing pattern. There seems at the very least to be some

tension between this and Dennett's insistence that we interpret agents according to normative criteria of rationality – that we see them as approximating to an ideal condition of rationality. How can we do this when our account of what it is to have a belief effectively mandates the attribution of conflicting and even contradictory beliefs to a single agent?

Dennett does have available to him a response to this line of argument. He would, I think, point out that constraints of rationality hold once we have committed ourselves to a particular pattern. We have to view agents as approximating to an ideal norm of rationality within the framework dictated by the pragmatic considerations that govern our choice of an explanation. We do not need to attribute conflicting beliefs and desires corresponding to the genuinely existing patterns. We just need to decide on one pattern and then attribute beliefs and desires accordingly.

This raises a deeper worry within Dennett's overall picture. How is Dennett to accommodate the causal dimension of the mental? Dennett denies the charge of epiphenomenalism (of effectively rendering the mental causally impotent), maintaining that his genuinely existing real patterns track causal phenomena.

> If one finds a predictive pattern of the sort just described one has *ipso facto* discovered a causal power – a difference in the world that makes a subsequent difference testable by the standard empirical methods of variable manipulation.
>
> (ibid., p. 43, n.21)

The notion of causation with which Dennett is operating appears to stress the counterfactual dimension of causation. He thinks that causation needs to be understood in terms of what *would* happen in suitably different circumstances, and this of course is precisely what one identifies when one identifies a pattern in an agent's behavior. There are, as we shall see further below in section 6.3, questions to be asked about whether the counterfactual account really does give us a suitably robust notion of causation. But even if we bracket these worries there appears to be a serious difficulty in Dennett's approach. There is no pragmatic element to causation. Causation is a metaphysical relation, not an explanatory relation. It makes sense to think about different explanations being required of a given event in different contexts. But nothing comparable holds for causation. If an event has a given cause, or set of causes, it can have no others. And yet Dennett has to maintain that there is indeterminacy in the causal origins of behavior. If an action can exemplify a range of different and incompatible patterns, and patterns effectively determine causation, then it follows that an action can have a range of different and incompatible causes. Few theorists who believe in the causal efficacy of the mental will accept this.

There are, then, two potential problems with Dennett's attempt to block the standard line of argument by appealing to real patterns. The first is a

general problem. Dennett argues with some plausibility that real patterns are not observer-dependent. Real patterns exist whether they are detected or not. But the actual criterion that he gives appears to allow the existence of incompatible patterns – patterns, for example, that generate conflicting predictions. In itself, this may not be a problem. There may be pragmatic reasons why one explanation might nevertheless be better (and therefore, perhaps, truer) than the other. But granting this generates a second, more specific problem. Dennett thinks that his real patterns are causal patterns (thus allowing his pattern explanations to be causal explanations), but then we seem to have the result that a single event can have a range of incompatible causes. This threatens to undermine the basic idea that commonsense psychological explanations can be causal explanations.

6.2 How anomalous is the mental?

In Chapter 3 we saw how Davidson's anomalous monism provides a powerful way of developing the picture of the autonomous mind. The key idea of anomalous monism is to drive a wedge between the relation of causation and the notion of explanation in a way that allows propositional attitudes to be causally efficacious without personal-level explanations being in any sense reducible to subpersonal explanations. Anomalous monism is a proposal to reconcile the following three basic principles.

1 *The principle of causal interaction.* There is causal interaction between the mental and the physical – as well, indeed, as causal interaction within the realm of the mental itself. Mental states are causally responsible for generating both behavior and other mental states.
2 *The principle of the nomological character of causation.* Causation requires strict causal laws. Part of what makes it the case that one event *E* causes another event *G* is the existence of a law linking events of the first type to events of the second type.
3 *The principle of the anomalism of the mental.* There are not, and cannot be, any strict causal laws holding over psychological states.

It is clear that the three principles are *prima facie* in conflict. It looks very much as if the first two principles jointly entail precisely what the third principle denies, namely, that there must be strict causal laws defined over mental states. As we saw in section 3.2, Davidson attempts to reconcile the three principles by arguing that mental causes are identical to physical events. Causal laws hold, not between events *per se*, but rather between events as described in certain ways. The principle of the anomalism of the mental, Davidson claims, holds only that there cannot be causal laws holding over mental events *when those events are characterized in psychological terms*. It is compatible with there being causal laws defined over mental events *when those events are characterized in physical terms*. It is the existence of these causal laws

holding over mental events under their physical description that preserves the nomological character of causality in the face of mental causation.

Since in this chapter we are concerned primarily with assessing the case for the picture of the autonomous mind, our interest is primarily in Davidson's arguments for the principle of the anomalism of the mental. Accepting the principle would have significant implications for how we view the relation between commonsense psychological explanation and lower-level approaches to the mind. It would mean, for example, that we would have to abandon the hope for a direct solution to the interface problem of the type proposed by the pictures of the functional and computational mind.

The essence of Davidson's argument is given in the following passage:

> There are no strict psychophysical laws because of the disparate commitments of the mental and physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual. There cannot be tight connections between the realms if each is to retain allegiance to its proper source of evidence ... The point is that when we use the concepts of belief, desire and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory.
>
> (Davidson 1970, pp. 222–223)

The argument can be broken down into two distinct steps. The first stage is the claim that the project of psychological understanding and psychological explanation is constitutively governed by considerations of rationality, coherence and consistency. We can only interpret people's behavior on the assumption that they are largely consistent and rational. When we find apparent inconsistencies and irrationalities, we have to rethink our attributions of desires and beliefs to restore consistency. Psychological understanding is seeing people's behavior as making sense in the light of what they want and the information they possess about the world. We cannot do this unless we view them as having largely consistent and coherent systems of belief and as doing what it is rational for them to do in the situation in which they find themselves.

The second step in the argument comes across rather more obliquely in the quoted passage. The principles and norms that govern the psychological realm are not, Davidson thinks, commensurable with the principles that govern the physical world. The norms of consistency and rationality cannot be understood in physical terms. They "have no echo", as Davidson puts it elsewhere, in the physical world. It is here, of course, that the real argument for anomalous monism is to be found, and it is unfortunate that Davidson's

position is not as clearly stated as it could be. In what exactly does the incommensurability consist? What has this incommensurability got to do with causation?

William Child has offered an appealing reconstruction of Davidson's argument in terms of what he terms the uncodifiability of rationality. Child starts off from two basic premises. The first is Davidson's central claim about personal-level psychological explanation, namely, that it is both governed and constituted by normative principles of rationality. The second premise is a claim about what would be the case if there actually were psychophysical laws of the type whose possibility Davidson is concerned to deny. If there actually were strict psychophysical laws then, Davidson and Child think, we would be in a position, at least in principle, to assimilate the behavior of people to the behavior of moving bodies. It seems clear, for example, that we can determine how the behavior of a moving body is prescribed by physical laws, and then use that information to make predictions about how a particular moving body will behave in a particular context – or to explain why it behaved the way it did in that context. The laws of physics, at least as they apply to ordinary-sized objects of the sort with which we interact on a daily basis, prescribe unique outcomes for physical interactions – or, if not unique outcomes, then at least outcomes with determinate probabilities. If there were psychophysical laws, then we would be in a position to map the principles of personal-level psychological explanation onto principles statable in the language of physics and then use these principles, in conjunction with physical descriptions of the relevant organisms, to determine unique outcomes for physical interactions involving persons. Psychophysical laws would allow us to treat people as physical systems and hence to predict their behavior in physical terms while still respecting the principles of personal-level psychological explanation. We would be able to switch more or less at will between the language of physics and the language of personal-level psychological explanation.

When we put these two premises together, it follows that the existence of psychophysical laws requires the existence of physically statable principles that will determine what it is rational for an agent to do or believe in a given context. There would have to be principles statable in a physical vocabulary that, in conjunction with a physical characterization of an agent in a particular context, would determine the rational course of action for that agent. These principles would allow us to identify in physical terms the mental states that should be attributed to an agent – and hence to get started on the process of psychological explanation/prediction.

The argument for anomalism that Child offers on Davidson's behalf is effectively that there can be no such physically statable principles – from which the anomalism of the mental follows by *modus tollens*. But why should we think that the principles of rationality cannot be stated in the language of science? Child offers a single line of argument with two strands. One strand has to do with practical rationality and the other with theoretical

rationality. The overarching claim is that the process of determining what it is rational to do (from either the first or third person perspectives) is *guided* by principles of rationality without being *dictated* by them. The process of actually applying the principles of rationality in a particular case is not rule-governed – and hence not the sort of thing that could be captured by any sort of overarching decision procedure, let alone one statable in a purely physical vocabulary. It might be clear which principles are relevant, but not how they should be brought to bear on the situation and weighed against each other. As Child puts the point, "in neither case [theoretical or practical] is there a fixed weighting or ordering of the competing considerations, or any definite rule for comparing them" (1993, p. 222).

In the case of practical reasoning, Child starts from the assumption that the question of what it is practically rational for a given person to do in a particular situation is the question of what that person should do in that situation. There is no gap, he thinks, between the requirements of practical rationality and the requirements of practical decision-making. With this very broad sense of practical rationality in mind, he argues that practical rationality cannot be reduced to means–end reasoning: "practical reasoning is not a matter of reasoning about the means to a predetermined set of ends, for it is not fixed in advance what the governing aim of a decision might be" (ibid., p. 222). Part of the process of practical decision-making is determining which of a range of competing ends is applicable in the situation in which one finds oneself, and this is not something that can be done mechanically or algorithmically. Ends are frequently conflicting and perhaps even incommensurable. There are no rules that will arbitrate between them.

Similar conflicts arise in theoretical reasoning. Here too we are confronted with a range of different factors that need to be weighed and compared. Some of these factors are epistemic and others not. We might distinguish, for example, between the purely epistemic issue of how propositions are logically or probabilistically related to each other from the more practical issue of what it would be rational for me to believe. There are cases where these two might come apart – it might be rational for me to have a belief that is neither entailed nor made more probable by my other beliefs. Even when we confine ourselves to the epistemic level, there are many competing desiderata. Consistency, fruitfulness, simplicity, explanatory power, broadness of scope – and so on. These different ideals might pull in different directions, and there is no decision procedure that will dictate how to resolve these conflicts.

For these reasons, then, Child argues on Davidson's behalf that there can be no principles formulable in a physical vocabulary that will determine what it is rational to do in a given situation – and hence, by *modus tollens*, that there can be no psychophysical laws. In fact, his arguments, if sound, will support the much stronger thesis that there are no statable principles at all that will determine what it is rational to do in a given situation. The difficulties posed by conflicting and competing considerations are difficulties for any attempt to formulate a rule-based conception of rationality. Of

course, one might respond that there is no need for a rule-based conception of rationality, but we can concede this requirement for the sake of argument. It seems clearly mandated by Davidson's conception of what a psychophysical law would have to look like. The interesting question is whether the requirement is too strong to meet.

In thinking about this one might begin by wondering whether Child and Davidson are not building too much into the notion of rationality. Child's understanding of practical rationality is a case in point. He explicitly states that "the characteristic question we face in practical rationality is what should I do in this particular set of circumstances" (ibid., p. 222). Once one has determined what it is rational to do in a particular situation, there is no further question to be asked about how one ought to act. One might well think, however, that if the scope of rationality is *that* broadly defined, then it is unsurprising that there will be no principles determining what it is rational to do in a given situation. After all, there are few people who think that we need to be able to find principles that will determine what would be the *right* thing to do in any given situation. Two questions naturally arise. The first is whether it is even legitimate to take the notion of rationality in this broad sense. There may be reasons for thinking that practical rationality cannot be understood in the very broad sense that Child proposes. The second question concerns the requirements of psychological explanation. When we make sense of people's behavior, in the service either of explanation or of prediction, do we really employ the broad notion of rationality that Child emphasizes? When we say that psychological explanation and prediction are governed by normative principles, do we really mean that we explain and predict behavior on the assumption that people will do what they *ought* to do, in some rich sense of 'ought'?

There is often a real question to be asked about whether one ought to do what it is rational to do. The prisoner's dilemma offers an interesting illustration. The prisoner's dilemma is a game (in the game theorist's sense, on which a game is a strategic interaction between two or more players) where the two players are prisoners being separately interrogated by a police chief investigating a crime who is convinced of their guilt, but as yet lacks evidence. He proposes to each of them that they betray the other, and explains the possible consequences. If both prisoners betray the other then they will both end up with a sentence of five years in prison. If both hold out and refuse to betray, then they will each be convicted of a lesser offence and both end up with a sentence of two years in prison. If either prisoner betrays the other without being implicated himself, however, then he will go free while the other receives ten years in prison. If we view each prisoner as motivated solely by the desire to minimize prison time, then it is clear that they will each rank the possible outcomes in interestingly different ways.[9] The best outcome for one prisoner (going free) will entail the worst option for the

9 The prisoner's dilemma is presented in a slightly different form in section 7.5.

other (ten years in prison). The second-best outcome for each prisoner is where neither betrays the other (two years in prison), while the third-best outcome is where both betray each other (five years in prison). So, if we represent prisoner 1's ordering of outcomes as $A > B > C > A$, prisoner 2 will have an ordering of the form $D > B > C > D$ – each prisoner's best-case scenario is the other's worst-case scenario, but they will rank the two scenarios where they do the same thing equally.

The abstract preference ordering at play in the prisoner's dilemma is exemplified in many social interactions (see section 7.5 for further discussion). In deciding what to do in this type of situation one is faced with a range of considerations. One of these considerations is the so-called dominance reasoning that reveals betraying the other player to be the rational strategy – rational on a sense of "rationality" that can indeed be given a very clear formulation in the basic principles of game theory and decision theory. The basic idea is that whatever player 2 does, player 1 is better off betraying her (if player 2 chooses to Betray, then player 1 will spend fewer years in prison if he also chooses Betrayal, while if player 2 chooses Hold Out, then player 1 is also better off choosing Betrayal). But other considerations may well come into play. One might not be motivated solely by the desire to minimize jail time. One might be motivated, for example, by a desire not to profit from someone else's misfortune. Or by a desire not to be a free-rider (a free-rider is someone who derives a benefit without paying the corresponding cost). The simple empirical fact that people who find themselves in prisoner's dilemma-type situations, both in the laboratory and in real life, frequently fail to take the dominant strategy shows that considerations such as these must come into play at least some of the time. The question, however, is how exactly to accommodate them.

There are two ways of proceeding. One might think that these additional considerations should be factored into how the game is described (into what game theorists call the game's pay-off table). This effectively changes what it would be rational to do. So, for example, for someone not prepared to be a free-rider it would simply be false that the most advantageous outcome in a prisoner's dilemma-type situation would be the one in which she betrayed while the other person held out. In fact, the outcome that seems most desirable on standard views of the prisoner's dilemma would become the least desirable if the disinclination to be a free-rider was factored into the pay-off table – and, as a consequence, defection would no longer be the dominant strategy. On the other hand, however, one might think that the person who refuses to take the dominant strategy because she is not prepared to be a free-rider is doing that *even though* the dominant strategy would have been the rational strategy. What she is doing is refusing to be rational. There is a way of thinking about the prisoner's dilemma on which the demands of rationality are just one of a range of considerations that might come into play – and hence on which deciding what it is rational to do does not automatically determine what one should do.

If we make this sort of distinction between what it is rational to do and what one should do, and allow that there is room for the two to come apart, then the very broad conception of rationality appealed to by Child begins to seem too broad. Determining what it would be rational for a person to do doesn't fix what that person should do – and conversely, the fact that there are no physically statable principles that will determine what a person should do in a particular situation cannot be taken as evidence that there are no physically statable principles of rationality. The argument for the uncodifiability of rationality starts to look somewhat weaker.

We seem to be back in the realm of competing considerations and different weights that Child and Davidson think are so inimical to the possibility of laws governing the mental. The issue is whether there might be laws governing how conflicts between, for example, ethical principles and the demands of rationality might be resolved. Do we have any reason to think that there could not be laws stating, for example, that the desire not to be a free-rider will trump considerations of instrumental rationality? If there were laws such as these then they could, in conjunction with a suitably codified conception of rationality, play the role in governing the psychological realm that the laws of physics play in governing the physical realm.

Things are not quite as simple as this, however. We need to take another look at the example. Suppose there is a true generalization that the desire not to be a free-rider will trump considerations of instrumental rationality in prisoner's dilemma-type situations. This would only be a genuine law if the two elements whose law-like connection is being suggested were genuinely independent of each other – that is to say, if there were a fact of the matter about what it is to desire not to be a free-rider other than not taking the dominant strategy when one is in a situation that has the structure of a prisoner's dilemma. But this assumption is begging the question against Davidson and those who think like him.

In order to see what is going on here it is useful to look at a passage from 'Psychology as philosophy' where Davidson explains how he came to appreciate the anomalism of the mental. Davidson's disillusionment with the idea that the mental might be a fixed law-governed system began with a set of experiments that he carried out when he was an experimental psychologist exploring rational decision-making. The issue in which he was interested was apparent breaches of the transitivity of preferences. It is, many have thought, a basic requirement of reason that one's preferences should be transitive – that is to say, if one prefers *a* over *b* and *b* over *c* then one should prefer *a* over *c*. Although this requirement is built into all standard versions of decision theory,[10] there is an empirical question about the extent to which ordinary reasoners respect this requirement and Davidson devised an experimental paradigm to explore this.

10 The best short introduction to choice theory I have encountered is Allingham (2002).

> Subjects made all possible pairwise choices within a small field of alternatives, and in a series of subsequent sessions, were offered the same set of options over and over. The alternatives were complex enough to mask the fact of repetition, so that subjects could not remember their previous choices, and pay-offs were deferred to the end of the experiment, so that there was no normal learning or conditioning. The choices for each session and each subject were then examined for inconsistencies – cases where someone had chosen *a* over *b*, *b* over *c* and *c* over *a*.
>
> (Davidson 1974, pp. 235–236)

Davidson discovered a curious pattern emerging as the sessions continued:

> It was found that over time intransitivities were gradually eliminated: after six sessions all subjects were close to being perfectly consistent ... If the choices of an individual over all trials were combined, on the assumption that his "real" preference was for the alternative of a pair he chose most often, then there were almost no inconsistencies at all. Apparently, from the start there were underlying and consistent values which were better and better realized in choice.
>
> (ibid.)

Here is how Davidson thinks that the experimental behavior supports the thesis of anomalous monism:

> The significance of the experiment is that it demonstrates how easy it is to interpret choice behaviour so as to give it a consistent and rational pattern. When we learn that apparent inconsistency fades with repetition but no learning, we are apt to count the inconsistency as merely apparent. When we learn that frequency of choice may be taken as evidence for an underlying consistent disposition, we may decide to write off what seem to be inconsistent choices as failures of perception or execution. My point is not merely that the data are open to more than one interpretation, although this is obviously true. My point is that if we are intelligibly to attribute attitudes and beliefs, then we are committed to finding, in the pattern of behaviour, belief and desire, a large degree of rationality and consistency.
>
> (ibid., p. 237)

There are two ways of looking at the initial apparent breaches of the transitivity of preferences. We can discount them as mere performance errors, looking through them to the underlying competence and rationality that is revealed over the entire length of the experiment. Or we can take them at face value, as breaches of transitivity that are corrected over time. But Davidson's point is not that we have no way of finding out what the fact of the matter is. It is the far more radical claim that there is no fact of the matter at all. Or rather, the

fact of the matter is determined by considerations that we bring to interpreting the situation – the considerations of rationality and consistency that Davidson identifies in the final sentence. There is, Davidson thinks, a very radical *indeterminacy* at the heart of the psychological, and it is this indeterminacy that is the real motivation for anomalous monism.[11]

One might wonder how this conception of indeterminacy can be reconciled with Davidson's insistence on the reality of mental causation. If a mental state such as a belief or a desire is identical to a neurophysiological state, then an indeterminacy in the realm of the psychological is presumably an indeterminacy in the realm of the neurophysiological, and it is hard to see how this should be understood. But we can put those metaphysical concerns to one side. Let us think instead about the example Davidson gives of indeterminacy. The point about the experiments he considers is that there are two different things that could be going on in the choice behavior. The subjects could either be starting out with inconsistent preferences and then gradually eliminating inconsistency – or they could be consistent all along, with their initial apparent inconsistency a function of performance errors. Davidson cannot see what would make one account true and the other false, except our deciding on one as a function of our own commitment to interpreting people's behavior so that they emerge as largely rational and consistent. His entire argument rests upon this being a metaphysical indeterminacy, rather than an epistemological indeterminacy. One way of assessing his argument would be to think about whether things really are as indeterminate as he takes them to be.

The key to the possibility of interpreting the choice behavior in different ways is that we have a choice between two different ways of understanding what a person's "real" preferences are. We can take their real preferences either synchronically or diachronically. That is to say, we can assume that what they choose at any given moment is what they really prefer, or we can assume that what they really prefer is what they choose most often. Imagine that a subject is confronted five times with a choice between *A* and *B* (suitably camouflaged, so that the subject doesn't realize that the same choice is being offered five times). The subject makes the following series of choices: *A*, *B*, *B*, *A*, *B*. According to the synchronic way of thinking about preferences, the subject doesn't have any stable preferences, but rather is oscillating between a preference for *A* and a preference for *B*. According to the diachronic way of thinking about preference, however, the person's real preference is for *B* over *A*, since she chooses *B* more frequently than she chooses *A*. Davidson's claim, as I understand it, is that the only reason to choose one or other of the synchronic or diachronic ways of thinking about a subject's real preferences is the need to interpret the subject's behavior so that it

11 It will be recognized that I am departing from standard ways of expounding anomalous monism here – anomalous monism is frequently taken to provide support for the thesis of the indeterminacy of the mental. I cannot see, however, how either thesis can be developed independently of the other.

comes out as consistent as possible. If consistency requires a synchronic interpretation, then so be it. Likewise for a diachronic interpretation.

But part of what makes the choice between the synchronic and diachronic interpretations seem so arbitrary is that we are given almost no information at all about the content of the choices being made. We are not told what the subjects were choosing between, nor about how much time separated the sessions. It is clear, however, that the details matter. Suppose that what is at stake is the choices that someone makes on five consecutive days in the university restaurant – with *A* being salad and *B* being steak and fries. Something like the synchronic approach seems far more appropriate here. It seems reasonable to think that the subject will have an identifiable preference on each day, and correspondingly unreasonable to look for an underlying "real" preference. As far as restaurants go, what people choose is generally what they want, and there is no need to look for consistency over time. There is no reason why, if I want salad on Monday, then I should want salad on Tuesday. But there are many situations, however, on which this does not hold. Let us suppose that I am faced with a series of choices in which what is at stake is whether or not to take a risk – with the risk being different each time. One choice might be, for example, between taking an exciting, challenging job in a different country, or a more familiar job closer to home – another might be between investing a lump sum on the stock market and placing the funds in a savings account. A third might be between taking up skydiving and starting to play squash. In each these situations let *A* be the risky activity (going abroad, investing in equities and taking up sky-diving) and let *B* be the less risky alternative. Here it does seem to make sense to ask whether I am a risk-taker, or whether I am risk-averse. It does make sense to ask whether there is a "real" preference underlying the individual choices I make. Moreover, it seems natural to try to identify this underlying preference by using something like the diachronic method. If I go for the safe option in the majority of cases then it looks very much as I am risk-averse. What we are trying to identify here is an underlying dispositional state – and the existence of an underlying dispositional state is inextricably tied to behavior over the long run.

Admittedly, Davidson's example of apparent indeterminacy in preferences is an illustration, not an argument. When we look at it in more detail, however, it becomes very unclear how representative it really is. Perhaps the indeterminacy that Davidson identifies is an artifact of the experimental paradigm – or even of the level of generality at which he characterizes that paradigm. We need something more to convince us that there is a genuine metaphysical indeterminacy here, rather than a degree of context-sensitivity in the procedures by which we might actually go about establishing preferences (or any other mental state) in a given context. In the last analysis, the issue here is really one of where the onus of proof lies. Since the anomalous monist is effectively arguing that an entire research project is doomed, it seems plausible to think that she needs to provide very strong reasons for

foreclosing on the possibility of psychophysical laws. The arguments in support of anomalous monism are at the very least debatable. They hardly compel assent.

6.3 The counterfactual approach

In the first two sections we have considered two ways of putting pressure on standard ways of thinking about the causal dimension of commonsense psychological explanation. In section 6.1 we explored Dennett's suggestion that psychological explanation is not causal explanation in anything like the standard way assumed by philosophers, namely, as depending upon the existence of causally efficacious inner items corresponding to the beliefs, desires and other propositional attitudes identified in psychological explanations. Psychological explanation is better viewed, Dennett suggests, as a matter of detecting real patterns in behavior and deploying those real patterns in the service of prediction and explanation. These real patterns are, Dennett insists, genuinely observer-independent. It is, contrary to his earlier instrumental understanding of psychological explanation, the existence of the real patterns that grounds the predictive/explanatory adequacy of commonsense psychology, rather than vice versa. Nonetheless, as we saw, Dennett's notion of a real pattern may well be not be strict enough to support the idea that psychological explanation is genuinely causal. He seems committed to the possibility that a given action might have a range of different and potentially incompatible causal antecedents.

The idea that there are causally efficacious inner items corresponding to beliefs and desires is closely bound up with a further element in the standard way of thinking about the causal dimension of psychological explanation. It is often assumed that *if* commonsense psychological explanations are causal explanations, then the generalizations of commonsense psychology must be strict causal laws. This assumption is essentially Davidson's principle of the nomological character of causation – the principle that all causal relations are law-governed. Section 6.2 explored the arguments for the anomalism of the mental – and in particular for the thesis that there can be no laws featuring psychological states (under psychological descriptions) – and hence, *a fortiori*, no causal laws holding over psychological states. The arguments for the anomalism of the mental offer one very powerful way of arguing for the autonomous picture of the mind. We saw, however, that there are reasons for being skeptical about the power of the arguments for the anomalism of the mental.

So where does this leave us? The standard way of thinking about the causal dimension of commonsense psychological explanation remains in play. We have not yet seen any conclusive reasons to deny that personal-level psychological explanation is a form of causal explanation, requiring the existence of causally efficacious internal items and dependent upon the existence of causal laws governing the relations between mental states and

between mental states and behavior. In the final section we explore a further challenge to the standard view. This is effectively a challenge to Davidson's principle of the nomological character of causation – to the idea that genuine causal explanation requires causal laws. It is the counterfactual approach to mental causation, an approach that promises to explain how there can be genuine causation without the existence of causal laws. The key to the counterfactual approach is the idea that a particular combination of mental states causally explains a given behavior if and only if it is true that in the absence of that combination of mental states the behavior in question would not have occurred – and, moreover, that that same combination of mental states would have led to the behavior in question even in different circumstances and background conditions. This is called the counterfactual theory because it makes the existence of causal relations dependent upon the truth of conditional statements that are counterfactual (that is, statements about what *would have* happened *if* the starting conditions had been different).

If this approach to mental causation can be made good, then it promises to go a considerable way towards dissolving the interface problem. If mental causation can be understood in counterfactual terms then we will not need to show how the causal generalizations of commonsense psychology can be subpersonally implemented (in the manner proposed by the functional picture of the mind). Nor will we need to explore how the subpersonal vehicles of personal-level psychological states can be causally efficacious in virtue of their structure in the manner proposed by the picture of the representational mind. Moreover, the counterfactual approach goes hand in hand with a downplaying of the significance of personal-level psychological generalizations (since these are no longer required to underwrite the causal dimension of commonsense psychological generalizations), there will be less scope for attacks on the theoretical poverty of commonsense psychology proposed by some proponents of the neurocomputational mind.

The counterfactual approach to mental causation can be developed either independently or as part of an overarching counterfactual theory of causation in general. For present purposes it will be easier to discuss it with reference to the special case of mental causation. Lynne Rudder Baker's theory of practical realism offers a clearly articulated version of this strategy – and, moreover, one that is explicitly targeted at philosophical orthodoxies about mental causation. Baker starts off from a counterfactual account of mental states:

> Whether a person S has a particular belief (individuated by a 'that–' clause in its attribution) is determined by what S does, says, and thinks, and what S would do, say and think in various circumstances, where "what S would do" may itself be specified intentionally. So, whether 'S believes that p' is true depends on there being relevant counterfactuals true of S. The antecedent of a relevant counterfactual may mention other

> of S's attitudes, but not, of course, the belief in question. If S is a speaker of a language, then the relevant counterfactuals concern her linguistic as well as her nonlinguistic behaviour. These counterfactuals bear the weight of revealing the "nature" of having beliefs and the other attitudes.
>
> (1995, pp. 154–155)

This aspect of practical realism bears some resemblance to philosophical behaviorism, as developed by Gilbert Ryle in *The Concept of Mind* and elsewhere. Rylean behaviorism also stresses the significance of conditional statements about what a person would do in particular situations, taking the truth of a mental state attribution (such as the attribution of a belief that *p*) to consist in the truth of certain conditionals about how that person would behave. Moreover, both Ryle and Baker take propositional attitudes to be properties of the whole person (as opposed to being discrete states or parts of the person). What makes it the case that a person believes that *p,* or desires that *q,* is that that person is disposed to behave in particular ways in particular circumstances. We do not need to talk about what is going on in their brain or central nervous system. There is no part of the person (a population of neurons, say, or a sentence in the language of thought) that can be identified with the belief that *p*.

However, there is one very important difference between the two positions. Ryle set out to offer a theory of mental states and, correlatively, a theory of psychological explanation. He was interested in explaining what mental states are and how citing mental states could provide useful explanations and predictions of behaviors. As far as he was concerned, however, a proper understanding of the mental reveals that psychological explanation should not be understood in causal terms at all. Baker, however, explicitly sets out to explain the causal dimension of psychological explanation. Practical realism is directed not against those who think that psychological explanations are causal explanations, but rather against those who think that psychological explanations can only be causal explanations if mental states are either identical to, or somehow realized in, discrete physical structures. Baker endorses, in a way that Ryle did not, a counterfactual theory of causation, linked to a counterfactual-based account of psychological explanation.

The basic idea of a counterfactual account of causation is that one event causes another in a particular set of circumstances if and only if, had the first event not occurred in those circumstances, the second would not have either and, were the first event to occur in similar circumstances, so too would the second. It is easy to see how a counterfactual account of causation can be developed to give an account of how psychological explanations can be genuinely causal. Effectively, what a psychological explanation offers is a complex event (that is to say, a combination of beliefs and desires in a particular set of circumstances) that stands in an explanatory relation to a particular action. An important part of what makes that relation explanatory is that it satisfies certain counterfactual constraints. Let us consider a

particular event of a certain type (say, a *G*-type event) occurring in a particular set of background conditions (which we can term *C*). Suppose we want to cite another event of a different type (say, an *F*-type event) as the cause of this event. What sort of connection are we looking for between the *F*-type event and the *G*-type event if we are to be convinced that the first really causally explains the second?

It seems plausible to follow Baker in thinking that there are two basic connections that must hold for there to be a causal explanatory connection between an *F*-type event and a *G*-type event. The first is that there should be a counterfactual dependence between the two events, such that had the *F*-type event not occurred in conditions *C*, the *G*-type event would not have occurred. A second connection might be that in any comparable situation in which an *F*-type event occurs, so too would a *G*-type event.[12] These two conditions are of course formulated in counterfactual terms. The central claim of Baker's practical realism is that the holding of these two conditions is a sufficient condition for there to be a causal connection between two events.

The counterfactual approach to psychological explanation offers a radical dissolution of many of the problems that we have been exploring. If the counterfactual approach is well grounded, then there is not really a problem of mental causation at all – at least in the sense standardly discussed by philosophers of mind and psychology. The counterfactual approach allows psychological explanations to be causal explanations without any need for causally efficacious internal items or causal laws. Much of the motivation for the functional and representational pictures of the mind disappears. The success of the counterfactual approach, however, depends upon taking the holding of certain counterfactuals to be constitutive of a causal explanation. All philosophers would agree that a genuine causal explanation implies certain counterfactual conditionals about what would happen were things to be otherwise – and, indeed, that the holding of these conditionals is what distinguishes a genuine causal explanation from a pseudo-explanation that trades on a mere coincidence. The real issue is whether this is all that there is to a genuine causal explanation.

One way of bringing the issues here into focus is to think about what would make a causal explanation true. The counterfactual theorist holds that there is nothing more to the truth of the causal explanation (and hence nothing more to the existence of a genuine causal relation) than the truth of the relevant counterfactuals about what would happen if circumstances were different. It is natural to think, however, that we cannot take the truth of counterfactuals as given. The truth of a counterfactual cannot be a brute fact, in the way that the truth of an ordinary assertoric statement can be a brute fact. The comparison is worth pursuing. It is natural to think that what makes my assertoric statement that the water is boiling true is the state of affairs of the water boiling. This state of affairs is the truth-maker for my

12 See Baker (1995, p. 122).

assertoric statement that the water is boiling. In this simple example, the truth-making relation is one of correspondence – correspondence between a statement and a state of affairs. Now consider a counterfactual statement to the effect that, had the stove not been switched on, the water would not have boiled. It seems clear that there is no corresponding state of affairs in the way that there is for the simple statement that the water is boiling, precisely because the statement is counterfactual – as things stand, the stove is on and the water is boiling. Nor, on the other hand, can we just take the counterfactual to be true without there being something that makes it true. So, what sort of truth-maker could there be for the counterfactual statement about what would have happened to the water had the stove not been switched on?

There is a basic choice to be made between two different ways of thinking about the truth-makers of counterfactual conditionals. On the one hand, one can think of counterfactuals as being made true by things that happen in the actual world. Whatever these things are, of course, they will not correspond to the relevant counterfactual statements in the straightforward way that the state of affairs of the water boiling serves as the truth-maker for the statement that the water is boiling. The truth-making relation will not be one of correspondence. Suppose, on the other hand, that one does want to think about the truth-makers of counterfactuals in terms of correspondence. Since counterfactuals cannot correspond to actual states of affairs, the correspondence must be to counterfactual states of affairs. Theorists who take this path generally deploy the notion of a possible world, the idea being that counterfactuals are made true by states of affairs in possible worlds that are suitably similar to the actual world. In a little more detail, a counterfactual conditional is true just if in the most similar possible world in which the antecedent is true, the consequent is also true. So, to return to the example of the water boiling, what makes it true that the water would not have boiled if the stove had not been switched on is that, in the nearest possible world in which the stove is not switched on, the water does not boil. The counterfactuals involved in psychological explanation work in exactly the same way. Suppose that I explain someone's switching the stove on in terms of their desire to boil an egg. This, according to the counterfactual approach, amounts to the following two claims. First, that had that person not desired to boil an egg she would not have switched the stove on. Second, that in any comparable situation in which she desired to boil an egg she would switch the stove on. Both counterfactuals are made true by what goes on in other possible worlds. The first counterfactual is made true by the fact that, in the nearest possible world in which she does not desire to boil an egg, she does not switch the stove on. The second counterfactual is made true by the fact that, in all nearby possible worlds in which she does desire to boil an egg, she switches the stove on.

Let us look at these two strategies in turn. Suppose we think that counterfactual conditionals are made true by what goes on in the actual world.

What features of the actual world could ground the truth of counterfactual conditionals? The only candidates seem to be laws governing the behavior of the objects featuring in the counterfactuals. Something that might make it the case that if the stove had not been switched on the water would not have boiled is that there is a law-like connection between water boiling and water being heated. This general law-like connection is underwritten by more specific laws governing the behavior of water in particular (such as the law that it boils at a certain temperature relative to atmospheric pressure) and the behavior of liquids and gases in general (the laws that explain how the application of heat brings about changes in temperature). This certainly gives us a way of understanding the truth of the counterfactual conditional. It achieves this by making the truth of counterfactuals a consequence of the holding of laws. What this means, however, is that the counterfactuals lose most of their explanatory power. Suppose we explain why the water boiled by citing the fact that it reached a temperature of 100 degrees Celsius (we can assume that we are at sea level in standard atmospheric conditions). There is indeed a true counterfactual associated with this, namely, that had the water not reached a temperature of 100 degrees Celsius (at sea level, in standard atmospheric conditions) it would not have boiled. But this is not what is really driving the explanation. What drives the explanation is the range of laws that govern the behavior of water and entail the counterfactual. It is the laws, rather than the counterfactual they entail, that is doing the explanatory work. This has clear implications for the strategy of appealing to counterfactual conditionals to dissolve the philosophical problems associated with mental causation. One of the aims of appealing to counterfactuals is to circumvent the need to appeal to causal laws to underwrite the project of psychological explanation. On this way of understanding the truth-makers for counterfactual conditionals, however, causal laws come back into the picture with a vengeance, eliminating one of the principal advantages claimed for the counterfactual approach – namely, the possibility of understanding causation without laws (and hence of rejecting Davidson's principle of the nomological character of causation).

What happens if we consider the second way of thinking about the truth-makers for counterfactuals? There are difficulties here also. Most of these difficulties emerge as soon as one asks what exactly possible worlds are. David Lewis, who pioneered counterfactual approaches to causation, is well known for having promoted a realist conception of possible worlds (Lewis 1986). Lewis's view is that there are indefinitely many genuinely existing possible worlds. Those possible worlds are concrete entities that have exactly the same degree of reality as the actual world. What makes the actual world actual is not that it possesses some form of real existence that no other possible world possesses. Rather, what makes the actual world actual is simply the fact that we inhabit it. Other possible worlds are no less actual (to their inhabitants) than the actual world is (to us). Now, if one is a realist about possible worlds in the way that Lewis is, then the machinery of possible

worlds provides a very clear way of understanding the truth-makers for counterfactual conditionals – and this, in fact, is one of Lewis's arguments in support of realism about possible worlds. There really are possible worlds and it is what goes on at those worlds that makes true counterfactual statements about what would happen if things were different. A counterfactual claim about this world is essentially an indicative claim about a counterfactual world, and it is true just if that counterfactual world is indeed as it is described as being. The truth-making relation is one of correspondence, just as it is with ordinary indicative statements about the actual world.

So, provided that one is a realist about possible worlds, the problem of identifying the truth-makers for counterfactual conditionals looks at the very least tractable – although there remain considerable difficulties in explaining both how we can have knowledge of possible worlds and how we should order possible worlds in terms of similarity. The problem, however, is that very few theorists have been prepared to follow Lewis in this realist approach to possible worlds. Most philosophers who deploy the notion of a possible world think of them as maximally consistent sets of propositions (Plantinga 1974), or as descriptions of ways in which things could have been (see the essays by Stalnaker collected in his 2003 collection). Many of these ersatz ways of thinking about possible worlds (as they have come to be known) seem to presuppose precisely the notion of possibility that they aim to explain. What is it for two or more sentences to be consistent, for example, other than for it to be possible for them simultaneously to be true? What is a way things could have been other than a possibility? Quite apart from these problems of potential circularity, however, there is a more fundamental problem directly germane to current concerns. We are looking for truth-makers for counterfactuals about, say, what would have happened had someone had different beliefs and different desires. These truth-makers must be sufficiently robust to motivate the idea that the relevant beliefs and desires are genuinely causally efficacious – and they must do this without being underwritten by causal laws governing how mental states relate to each other and to behavior. It is far from clear, however, that any of the ersatz ways of thinking about possible worlds could offer sufficiently robust truth-makers to allow us to dispense with causal laws. We do not have any indication of why certain sentences are consistent with each other, or the world might have been *this* way rather than *that* way.

There is, therefore, a very real challenge for proponents of the counterfactual approach to mental causation and psychological explanation. The counterfactual theorist is committed to a very strong understanding of counterfactuals, since the truth of appropriate counterfactuals is all that grounds the truth of the causal statements featuring in psychological explanation. This in turn raises the question of how we should understand the truth-makers for these counterfactuals. One very natural way of understanding the truth of counterfactuals is in terms of causal laws. The thought here is that it is laws about what must be the case that explain what would be the

case were circumstances different. This way of grounding counterfactuals is of course barred to the counterfactual theorist, who is seeking to explain how there can be causation without laws. The most plausible resource for the counterfactual theorist appears to be some form of possible worlds theory of counterfactuals, but the only version of possible worlds theory (namely, realism about possible worlds) that would uncontrovertibly do the job would strike many theorists as unpalatable. The challenge, therefore, for the counterfactual approach to mental causation is to explain what makes counterfactuals true without appealing either to causal laws or to concretely existing possible worlds. It is far from clear that this will be easily achieved.

6.4 Overview

According to autonomy theorists, the interface problem as I have presented it is ill-defined because it rests upon a misunderstanding about the nature of commonsense psychological explanation, and in particular about what it takes for commonsense psychological explanation to be a form of causal explanation. As far as the autonomy theory is concerned, the problem comes with the two requirements placed upon causal explanation by the standard way of thinking about mental causation. The first is the requirement that commonsense psychological explanations can only be causal if there are causally efficacious inner physical items corresponding to the psychological states that they identify. The second is the requirement that there be causal laws defined over commonsense psychological states. Accepting the conception that these requirements impose on what it would be for commonsense psychological explanation to be causal is effectively to treat commonsense psychological explanation as on a par with the various different types of explanation operative on the subpersonal level – as engaged in the business of identifying causes and causal laws. Once we see commonsense psychological explanation as engaged in effectively the same type of explanatory project as, say, cognitive neuroscience, then it clearly becomes imperative to ask how the respective projects link up with each other. This takes us to the interface problem as formulated in Chapter 2. Rejecting one or both of the two requirements, however, allows us to identify an incommensurability between personal-level and subpersonal-level explanations – and hence to suggest that the interface problem is ill posed.

In this chapter we have been exploring different ways of challenging these two key requirements. Dennett's conception of real patterns is offered in opposition to both requirements. According to Dennett, commonsense psychological explanation can count as a form of causal explanation without satisfying either of the two requirements because it rests upon identifying real patterns in the behavior of agents and intentional systems. These real patterns hold at the level of the whole system. They are weaker than causal laws and they do not depend upon the existence of causally efficacious internal items. Davidson's anomalous monism, in contrast, accepts the first require-

ment without the second. There do indeed have to be causally efficacious internal items – and, moreover, these causally efficacious internal items have to be identical to the psychological states that feature in the commonsense psychological explanations. However, for the reasons discussed in section 6.2, the anomalous monist argues that there are no strict causal laws defined over those causally efficacious inner items *when they are characterized in psychological terms* (although there are causal laws defined over them when they are characterized in physical terms). The final challenge explored (in section 6.3) rejects both requirements with the claim that there is nothing more to the truth of causal psychological explanation than the truth of certain counterfactual statements about what would have occurred had the agent had different beliefs and desires, or had the circumstances been relevantly different.

It has emerged in this chapter that there are serious difficulties for each of these challenges. The real patterns approach proposed by Dennett seems to be too generous in how it counts real patterns. It is committed to existence of real patterns that are incompatible – and hence to the possibility that a single event might have a range of incompatible causes. Davidson's anomalous monism faces a different problem. The argument for the anomalism of the mental (as I have reconstructed it) depends upon accepting a radical metaphysical indeterminacy in the realm of the psychological, and a concomitant incommensurability between personal and subpersonal levels of explanation, that has yet to be established. We have not been given, many philosophers will feel, a strong enough case for abandoning an entire research program in cognitive science and empirical psychology. The problem with the counterfactual approach, in contrast, is that it seems very difficult to provide a sufficiently robust account of what makes counterfactuals true if one deprives oneself of the resources offered by causal laws.

It is, of course, far too early to say whether any of these difficulties are insuperable. It should be clear, however, that they are serious and, indeed, serious enough to prevent us from foreclosing on the interface problem in the manner proposed by proponents of the picture of the autonomous mind. In the remainder of this book I will adopt the working assumption that personal and subpersonal levels of explanation are not as radically incommensurable as the autonomy picture maintains, and hence that the interface problem remains in play.

7 The scope of commonsense psychology

- Thinking about the scope of commonsense psychology
- Implicit and explicit commonsense psychology
- Modest revisionism
- Narrowing the scope of commonsense psychology (1)
- Narrowing the scope of commonsense psychology (2)
- A suggestion?

The previous chapter explored some ways of thinking about psychological explanation and the interface problem associated with the picture of the autonomous mind. In the approaches of Dennett, Davidson and those who offer a deflationary account of mental causation in terms of counterfactuals we find different attempts to reconfigure what one can think of as the standard conception of psychological explanation. Part of the aim of the autonomous picture of the mind is to show that the interface problem should not be taken seriously. Personal-level commonsense psychological explanation can be understood on its own terms and does not require validation from subpersonal levels of explanation. In fact, there can be no such validation, due to the radical incommensurability between personal and subpersonal levels of explanation. If the autonomy picture is well grounded, then the interface problem ceases to be a pressing concern. Let us suppose, however, that the proponents of the autonomous mind have yet to make their case, so that the standard conception of psychological explanation remains in play. This leaves us with the interface problem as originally presented in Chapter 2 – with the obligation to explain how the personal-level explanations of commonsense mesh with the explanations given at levels of explanation lower down in the hierarchy of explanation.

This chapter is devoted to a more general question that sets the framework for the interface problem and that in an important sense determines its significance. The question is one about the scope of commonsense psychology. How central a role does commonsense psychological explanation play in our understanding of ourselves and others? How significant is it in allowing us to interact socially? How widespread is the practice of commonsense psychological explanation? How deeply embedded is it in our everyday social practices and interactions. One's view of the importance of the interface problem will be a direct function of how one responds to these questions – as indeed do the resources one has to deal with it. The more central com-

monsense psychology turns out to be, the more important it is to resolve the interface problem – and the less likely it is that we will be able to do so by appeal to a single phenomenon (such as language, for example).

In section 7.1 I sketch out different ways of thinking about the scope of commonsense psychology. At one extreme is the broad construal of the scope commonsense psychology, which sees it as guiding all our social interactions – either explicitly or implicitly. At the other is the narrow construal, according to which we only ever make explicit use of commonsense psychology and we should be wary of attributing implicit knowledge of commonsense psychology. The real issue, it appears, is how we characterize those instances of unreflective social understanding where we are not consciously and explicitly making use of commonsense psychology. According to the broad construal, in such situations we are making implicit use of commonsense psychology. Section 7.2 considers how we might understand commonsense psychology as an implicit theory, comparing the thesis that we have implicit or tacit knowledge of commonsense psychology with the thesis that we have implicit or tacit knowledge of the syntactic principles of a language. In section 7.3 we consider an alternative approach, which sees the processes of explanation and prediction as involving projections of ourselves into other people's situations and using our own mind as a model of theirs in order to work out what to do. There are two versions of this *simulationist* proposal, one of which gives a way of thinking about unreflective social understanding that need not always involve the machinery of propositional attitude psychology. In the next two sections we consider ways of putting flesh on the bones of the narrow construal. Section 7.4 offers some general reasons for thinking that the scope of commonsense psychology might not be as dominant as it tends to be taken to be, while section 7.5 explores ways of understanding a range of social interactions and social situations that do not involve commonsense psychology.

7.1 Thinking about the scope of commonsense psychology

There are two different ways of thinking about the scope of the conceptual framework of propositional attitude psychology. One might, first, think of propositional attitude psychology as a *privileged* level of explanation, on the grounds that using the tools of propositional attitude psychology to explain and/or predict behavior allows us to capture commonalities and patterns in thought and action that cannot be captured at lower levels of explanation (e.g. Fodor 1987). One might, second, think of propositional attitude psychology as a *dominant* level of explanation. Whereas the notion of privilege is *qualitative*, the notion of dominance is *quantitative*. The dominance claim is one about how we actually go about explaining and/or predicting the behavior of other thinking subjects. Effectively, it is the claim that when we need to understand other people as psychological subjects and genuine agents, we

(as a matter of fact) almost invariably use the explanatory framework of propositional attitude psychology.

These two ways of thinking about the scope of commonsense psychology do not necessarily go together. One can think that commonsense psychology provides a privileged level of explanation, without thinking that it is dominant. One might, for example, think that commonsense psychology is too slow and computationally demanding to be used in many social situations. And one can equally think that commonsense psychology is our dominant tool for making sense of ourselves and others without thinking that it is privileged. The second of these two positions has several distinguished exponents. Paul Churchland and other eliminative materialists have suggested that commonsense psychology will eventually be replaced by a theory capable of dealing with complexities that propositional attitude psychology cannot tackle – a theory that will be derived from neuroscience rather than from commonsense psychological concepts. They do not doubt that commonsense psychology is currently our dominant tool for interpersonal cognition. What they dispute is that it is in any sense privileged. According to Churchland, we are in the unfortunate position of having to use a theoretical framework that is (in his opinion) demonstrably flawed, limited and stagnant – and we are condemned to remain in that position until our scientific understanding of the brain has made advances that can now barely be contemplated. A broadly similar conclusion has been reached by Stephen Stich, who also argues forcefully against the allegedly privileged status of propositional attitude psychology by attacking the notion of content upon which propositional attitude psychology rests (Stich 1983). Unlike Churchland, Stich does not look to completed neuroscience to provide the privileged level of explaining and predicting behavior. Rather, he offers the prospect of a purely syntactic version of the computational theory of mind in which the notion of content has no place. The syntactic theory of mind shares with Churchland's eliminative materialism, however, the thesis that the content-involving notions of propositional attitude psychology currently dominate our explanatory and predictive practices. The syntactic theory of mind is a promise for the future, not a description of the present.

These two positions apart, however, the privileged nature of commonsense psychological explanation is almost universally accepted. But what exactly is involved in taking commonsense psychology to be our dominant (as opposed to privileged) way of understanding ourselves and others? It is clear that we at times make explicit use of commonsense psychology. Sometimes we work forwards from what we know of someone's beliefs and desires to what we think they will do. Sometimes we work backwards from their behavior and general knowledge of how their minds work to their particular motivations for acting in a certain way.

But introspection *seems* to suggest that such explicit use of commonsense psychology is relatively infrequent. We spend most of our lives negotiating

our way through the social world, adapting our behavior to that of other people, taking part in joint activities, and so on. And we are in fact remarkably good at it. We navigate the social world with no less skill and dexterity than we manifest in navigating the physical world. But only a very small fraction of the time do we seem to make *explicit* use of commonsense psychology. It is relatively infrequently that we explicitly attribute propositional attitudes to other agents and then use those attributed attitudes to explain their behavior. It is natural, then, to ask what we are doing the rest of the time. What underwrites our skills in social understanding and social coordination on those occasions when we are not *explicitly* deploying the categories and tools of commonsense psychology?

This question is not asked as frequently as it might be because there is an equivocation in the expression of commonsense psychology – and indeed in the various expressions used interchangeably with it, such as theory of mind, folk psychology and propositional attitude psychology. On the one hand, the terms are used descriptively to characterize the complex of social abilities and skills possessed by all normal, encultured, non-autistic and non-brain-damaged human beings. In this rather weak sense it is trivially true to say that all our social interactions are governed by commonsense psychology. This is to say nothing more than that we use our skills in social understanding and social coordination in our social interactions. But the notion of commonsense psychology is also used in a much less neutral way; to characterize what is in effect a particular conceptual framework deemed to govern our social understanding and social skills. Here is a useful characterization of this second way of thinking about commonsense psychology from the introduction to a collection of important essays on commonsense psychology:

> It has become a standard assumption in philosophy and psychology that normal adult human beings have a rich conceptual repertoire which they deploy to explain, predict and describe the actions of one another and, perhaps, members of closely related species also. As is usual, we shall speak of this rich, conceptual repertoire as 'folk psychology' and of its deployment as 'folk psychological practice'. The conceptual repertoire constituting folk psychology includes, predominantly, the concepts of belief and desire and their kin – intention, hope, fear, and the rest – the so-called propositional attitudes.
>
> (Davies and Stone 1995a, p. 2)

This is a general characterization designed to leave room for more determinate theories about how exactly the concepts of the propositional attitudes are applied in commonsense psychological explanation. So, there are really three different ways of thinking about commonsense psychology.

1 The complex of skills and abilities that underlie our capacities for social understanding and social coordination.

2 A particular conceptual framework for social understanding and social coordination based upon the propositional attitudes.
3 A particular way of applying the conceptual framework in (2) in the service of explanation/prediction.

The important distinction at the moment is between the first and second ways of thinking about commonsense psychology. We will return to different ways of thinking about how the conceptual framework of commonsense psychology might be applied in section 7.3.

There is a danger in not keeping these different ways of thinking about commonsense psychology clearly distinct. It is obviously true that we are constantly using our skills in social understanding and social coordination, but far less obviously true that we are constantly applying a conceptual framework based upon the propositional attitudes. There is a question here that it is important to keep open. Granted that we sometimes do make reflective and explicit use of the concepts of commonsense psychology in making sense of the behavior of others, should we conclude that our *unreflective* social understanding involves an implicit application of the concepts of commonsense psychology in the interests of explanation and prediction? Should we conclude that all our social understanding involves deploying the concepts and explanatory/predictive practices of commonsense psychology, even when we are not aware of doing so?

We can distinguish two conceptions of the scope of commonsense psychology – or, more accurately, two ends of a spectrum of conceptions of the scope of commonsense psychology. At one end lies the narrow construal of the domain of commonsense psychology. According to the narrow construal, the domain of commonsense psychology should not be presumed to extend further than those occasions on which we explicitly and consciously deploy the concepts of commonsense psychology in the services of explanation and/or prediction. At the other end of the spectrum lies the broad construal, which makes all social understanding a matter of the attribution of mental states and the deployment of those attributed states to explain and predict behavior, whether that is what we are aware of doing or not.

Most philosophers adopt some version of the broad construal of commonsense psychology. The broad construal of commonsense psychology fits in with a particular way of interpreting the distinction between personal and subpersonal explanation. It is only a short step from the idea that all social understanding and social coordination involves applying the categories of commonsense psychology to the idea that we can only think about agency at the personal level in terms of the conceptual framework of commonsense psychology, so that the domain of the personal level becomes co-extensive with the domain of the propositional attitudes.

One reason for the widespread acceptance of something like the broad construal of commonsense psychology is that philosophers do not have many alternative models of how behavior might be understood at the personal

level. Philosophers of mind and action tend to operate with a clear-cut distinction between two ways of understanding behavior. We can either understand behavior in intentional terms, as rationalized by propositional attitudes, or in non-intentional terms. It is standard to distinguish, for example, between an arm-raising that is intentional, comprehensible as issuing from a particular nexus of beliefs and desires, and one that is the result of a reflex response, or of someone else lifting my arm for me. It seems clear that social understanding does not involve understanding the behavior of others in either of these latter two ways. So, if the choice really is a stark one between taking behavior to be unintentional in one of these senses, on the one hand, and taking it to be intentional in the sense of being rationalized by propositional attitudes on the other, then it is easy to see why unreflective social understanding should be widely thought to involve the tacit application of commonsense psychology.

Yet the interface problem does not depend upon the broad construal of commonsense psychology. The interface problem is the problem of explaining how commonsense psychological explanations interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy – and this problem arises however one construes the domain of commonsense psychology. When the distinction between personal and subpersonal levels of explanation was introduced in Chapter 2, commonsense psychological explanation was put forward as a paradigm example of personal-level explanation, but not as our only way of thinking about personal- level explanation. The possibility of ways of explaining behavior and interacting with other psychological subjects at the personal level that do not involve applying the conceptual framework of commonsense psychology remains very much open.

The issue is very important for how we think about the significance of the interface problem. Many philosophers have thought, for example, that certain features of commonsense psychological explanation will prove particularly difficult to understand in subpersonal terms. We saw a clear example of this type of thinking when we looked at the picture of the autonomous mind, and in particular at the arguments that the norms of rationality and consistency that govern propositional attitude explanation cannot be understood in terms of the causal generalizations operative at the subpersonal level. The significance of those worries within the overall project of providing a satisfactory account of the mind is directly correlated with how one construes the scope of commonsense psychological explanation. The broader the scope accorded to commonsense psychological explanation the more pressing the problem will be. Conversely, the narrower the scope of commonsense psychological explanation the more circumscribed the problem will be.

In fact, if something like the narrow construal of the scope of commonsense psychology turns out to be true, it may be that the personal-level

mechanisms that we use much of the time to navigate the social world do not present any of the difficulties that seem to make the interface problem so intractable. They may depend upon mechanisms that can straightforwardly be identified and understood at the subpersonal level. Another possible consequence of the narrow construal is that it may open up ways of dealing with the interface problem that would not be available if the commonsense psychology were as dominant as the broad construal takes it to be. Suppose, for example, that we only ever deploy the conceptual framework of propositional attitude psychology on those relatively infrequent occasions when we consciously and explicitly reflect on why a person has acted a certain way, or on how a person will behave. It may turn out that the key to explaining what is going on has to do, not so much with the psychological dimension, but rather with the fact that conscious and explicit reflection is going on. It might be, for example, that such conscious and explicit reflection always takes a linguistic form, and that an account of commonsense psychology will emerge from a more general account of linguistic thought.

7.2 Implicit and explicit commonsense psychology: the broad construal

The concept of commonsense psychology is called upon to do a number of different jobs. It is important to distinguish them. In most general terms we can describe commonsense psychology as a set of very basic skills – skills that allow us to navigate through the social world and to accommodate ourselves to the behavior of others. "Commonsense psychology" in this sense simply denotes a manifest set of abilities. An analogy that springs to mind is with comparable sets of basic skills and abilities in other domains. Psychologists, anthropologists and computer scientists have developed the idea that we possess a naïve physics that allows us to navigate through the physical world – to discriminate different types of material objects, fluids and varieties of "stuff" in ways that underwrite certain expectations about how they will behave and that allow us to manipulate them.[1]

But theorists in many different areas also frequently use the concept of commonsense psychology to characterize a set of generalizations about human behavior and its motivation that are often taken to be platitudes or truisms. For analytical functionalists, for example, these platitudinous generalizations define the functional roles of the intentional states featuring in commonsense psychology, and the total set of such generalizations forms a theory. This use of "commonsense psychology" to refer to a more or less theory-like structure is common to eliminativist neurophilosophy and the representational theory of mind, as well as to various strands in the autonomy approach.

1 Patrick Hayes has provided influential statements of the significance of naïve physics for artificial intelligence and robotics. See Hayes (1985a, 1985b), together with the other papers collected in Hobbs and Moore (1985). A brief overview will be found in Proffitt (1999).

Analytical functionalists stress the implicit nature of the theory and suggest that the generalizations of the theory can be made explicit by a process of compiling platitudes (see the passage from Lewis quoted on p. 59 above). Paul Churchland takes a broadly similar view of how we might go about discovering the generalizations of commonsense psychology:

> A thorough perusal of the explanatory factors that typically appear in our commonsense explanations of our internal states and our overt behavior sustains the quick "reconstruction" of a large number of universally quantified conditional statements, conditions with the conjunction of the relevant explanatory factors as the antecedent and the relevant explanandum as the consequent. It is these universal statements that are supposed to constitute the "laws" of folk psychology.
>
> (1981, pp. 52–53)

The same basic view of commonsense psychology is at work in the representational theory of mind. Fodor's attitude to commonsense psychology is more guarded than either of the two yet considered:

> An explicit psychology that vindicates commonsense belief-desire explanations must permit the assignment of content to causally efficacious mental states and must recognize behavioural explanations in which covering generalizations refer to (or quantify over) the contents of the mental states that they subsume … I don't, however, have a shopping list of commonsense generalizations that must be honoured by a theory if it wants to be ontologically committed to bona fide propositional attitudes. A lot of what commonsense believes about the attitudes must surely be false (a lot of what commonsense believes about *anything* must surely be false) … On the other hand, there is a lot of commonsense psychology that we have – so far at least – no reason to doubt and that friends of the attitudes would hate to abandon. So, it's hard to imagine a psychology of action that is committed to the attitudes but doesn't acknowledge some such causal relations among beliefs, desires and behavioural intentions (the 'maxims' of acts) as decision theories explicate.
>
> (1987, pp. 14–15)

Although Fodor leaves open the possibility that much of commonsense psychology may well be mistaken and ultimately corrected by some or other part of scientific psychology, he still holds that we learn commonsense psychology "at our mother's knee". Consequently it is not hard to discover – even though a completed cognitive science may not vindicate all that we discover.

Proponents of the autonomous mind would question the supposed theory-like nature of commonsense psychology, challenging in particular the view that it is continuous with predictive scientific theories. And, as we have

seen, it is characteristic of autonomy theorists to deny that commonsense psychological explanation is simply a matter of subsuming behavior under causal explanatory generalizations. Nonetheless, even Davidson is prepared to accept that commonsense psychological explanation employs generalizations about the behavior of rational agents. He describes them as a form of "practical wisdom" that allows us to impose a rational pattern on behavior (1970, p. 219). Similarly, McDowell's conception of "a style of explanation in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be" (1985, p. 389) presupposes the existence of a set of normative principles (perhaps not precisely codifiable) that determine what a rational agent ought to do in certain situations, given certain beliefs and desires.

All these authors share a basic assumption about the relation between reflective and unreflective commonsense psychology. This is the assumption that the set of commonsense psychological principles to which most of us would unhesitatingly assent forms part of a larger body of tacitly known principles that guide our social behavior and social understanding in those situations where we are not explicitly deploying the concepts and tools of propositional attitude psychology. The assumption emerges very clearly in the following passage from David Braddon-Mitchell and Frank Jackson:

> Trees and planets behave in relatively regular ways. When the wind blows a tree moves in much the same way each time. Mars moves through the sky in a highly predictable way. By contrast, human beings move in a quite bewildering variety of ways. Nevertheless we often succeed in predicting what they will do. How do we do this? By treating them as subjects with mental states. By observing what they do and say, we arrive at views about what they are thinking, what they desire and closely associated views about their characters, mental capacities and in general about their psychological profiles. We then, in terms of these profiles, predict what they will do. We have, then, great facility in moving backwards and forwards from behavior in situations to mental states. Think of what is involved in playing a game of tennis, crossing a road at traffic lights or organizing a conference. The antecedent probability that Jones will move her body in such a way that the ball will land where you have most trouble retrieving it, or that drivers will move their bodies in such a way that their cars will stop when the light turns red, or that a number of human bodies will move from various corners of the globe to end up at the same time in one conference centre, is fantastically small. Yet we make such predictions successfully all the time ... The fact that we can make the predictions shows that we have cottoned on to the crucial regularities – otherwise our predictive capacities would be a miracle. They show that we have an implicit mastery of a detailed, complex scheme that interconnects inputs, outputs and mental states.
>
> (1996, pp. 56–57)

On this view, the concepts and generalizations that we deploy when we explicitly try to explain or predict the behavior of others in terms of propositional attitude concepts are really just the tip of the iceberg – a small part of a vastly more complicated conceptual framework that governs all our social interactions and social understanding.

It should be clear, once this working assumption is brought into the open, that it is an empirical hypothesis about the psychology of social understanding – about the psychological mechanisms that people employ to understand themselves and others. It should also be clear that it is an empirical hypothesis that brings with it a considerable theoretical commitment, namely, to explain the nature of our *implicit knowledge* of commonsense psychology. The very idea of implicit knowledge is rather obscure and although the notion is widely deployed in psychology, cognitive science and linguistics, there is no accepted and worked out theory that can be unproblematically applied to the case of commonsense psychology.

The central case to which the notion of implicit knowledge has been applied is our understanding of the syntactic structure of our language (Chomsky 1980, particularly Chapter 3, and Miller 1997, for a philosophical overview). But there seem to be significant disanalogies between implicit syntactic understanding and implicit commonsense psychological understanding. Whereas we are told by linguists that the rules of syntax are relatively precise, the generalizations of commonsense psychology seem to hold for the most part (*ceteris paribus* – all other things being equal). Whereas the rules of syntax are hierarchically structured in a way that determines which rule is to take precedence in a given situation, the generalizations of commonsense psychology (as they are most frequently understood) throw up different and competing explanations or predictions of a given behavior. It is a highly context-sensitive matter to determine which commonsense psychological generalizations might be applied in a given situation, far more so than it is with syntactic principles. So, although some philosophers have offered theories of how we might understand the implicit knowledge that seems to be implicated in linguistic understanding (Evans 1981; Peacocke 1989; Davies 1989), these disanalogies stand in the way of applying those theories to our implicit knowledge of commonsense psychology.

We can put the point in terms of the distinction between modular and non-modular cognitive processes discussed in section 2.1. Some cognitive processes are open-ended and involve bringing a wide range of information to bear on very general problems. These are the non-modular processes, in contrast to lower-level, modular cognitive processes that work quickly to provide rapid solutions to highly determinate problems (Fodor 1983). Modular processes have certain characteristic features. They are *domain-specific*, applying only to a relatively circumscribed range of situations. They respond automatically to stimuli of the appropriate type (*mandatory application*) and they are unaffected by other types of cognitive processing (*informational encapsulation*). The types of processing involved in understanding the

syntactic structure of sentences are paradigmatically modular. But this is not the case for the implicit knowledge that is being claimed of commonsense psychology. It is hard to see how one might demarcate the field of social situations and identify the relevant stimuli that will trigger the operation of the "commonsense psychology module".[2] Nor is commonsense psychological understanding insulated from other types of cognitive processing. In explaining and predicting the behavior of other people we make use of all the collateral information we can get hold of. Psychological understanding is profoundly context-sensitive and the context is not purely social.

It is true that it is common practice among philosophers and psychologists engaged in debates about the psychology of social understanding and the nature of "theory of mind" to treat commonsense psychology as modular (see, for example, Leslie 1991 and Baron-Cohen 1995). Yet the precise relation between what Leslie terms the Theory of Mind Mechanism and the type of modules discussed by Fodor has received relatively little attention (but see Segal 1996, for a useful discussion). Although some have argued that we can treat commonsense psychology as a Fodorean module (e.g. Scholl and Leslie 1999), this seems a hard case to make. There may well be a coherent sense in which commonsense psychology is modular, but it is unlikely to be modular in the Fodorean sense. But our existing theories of implicit or tacit knowledge have been developed with Fodorean modules in mind. It is not obvious how they might transfer over to the case of commonsense psychology.

Michael Dummett has offered a very different conception of implicit knowledge that is directed at personal-level skills and abilities, rather than at subpersonal modules (Dummett 1993).[3] He starts off from the obvious question any account of tacit or implicit knowledge has to confront. What makes it the case that one proposition or set of propositions is implicitly known rather than another? It seems clear that there could be many different implicitly knowable bodies of knowledge that could equally account for any given ability. What would make it the case that a particular one of these was *the* implicitly known body of knowledge underlying the ability in question? Note that this is not a problem about how we could identify the relevant theory. The question is metaphysical rather than epistemological. It has to do with what would make it the case that one theory is tacitly known rather than all the others that could be tacitly known – not with how we could go about working out which theory that is.[4]

2 See Fodor (2000) for an elegant argument that this is impossible. This argument is presented and discussed in section 8.4.

3 Dummett himself thinks that linguistic understanding falls into this category. However, we need not follow him in this. We can leave open the possibility of a modular account of linguistic understanding, while taking Dummett's views seriously as a possible account of implicit knowledge of commonsense psychology.

4 Miller (1997) is a general introduction to current theories of tacit knowledge. The papers by Crispin Wright and Gareth Evans in Holtzman and Leich (1981) are difficult but well worth reading. See also Davies (1986, 1989).

In Dummett's view, it is only correct to talk about a particular proposition being tacitly known if the subject to whom such knowledge is ascribed can "acknowledge as correct a formulation of that which is known when it is presented" (1993, p. 96). There is a certain plausibility in thinking about unreflective commonsense psychology in these terms. Most of us are instantly prepared to acknowledge the truth of various commonsense psychological generalizations when they are explicitly formulated – and it is of course upon this that theorists such as Lewis are trading on when they suggest that we might come to a complete formulation of the principles governing commonsense psychological understanding by collating all the platitudes about mental states and their interactions that receive widespread acceptance.

Dummett asks us to imagine someone who has learnt to play chess simply by having his errors corrected, without having ever explicitly come across explicit formulations of any of the rules of the game. He comments:

> It would be unthinkable that, having learnt to obey the rules of chess, he should not then be able and willing to acknowledge those rules as correct when they were put to him, for example, to agree, perhaps after a little reflection, that only the knight could leap over another piece. Someone who had learned the game in this way could properly be said to know the rules *implicitly*. We might put the point by saying that he does not merely follow the rules, without knowing what he is doing: he is guided by them.
>
> (ibid., p. 96)

The analogy here may be too crude, however. Dummett's claim about the knight's move is plausible enough, but does not obviously carry over to more complex and theoretically interesting cases of implicit knowledge. Suppose we consider not the imaginary subject's implicit mastery of the basic rules of chess, but rather his mastery of certain basic principles of chess strategy – say, that the aim of the opening is to gain control of the four central squares or that one shouldn't launch an attack before castling. A subject can perfectly well be "guided" (in whatever sense being guided differs from following) by such principles even though he would vehemently deny their truth were he to encounter them in a book on chess. And this is why, of course, grandmasters are not always the best authorities on the games they have played. On the occasions when we explicitly employ the principles of reflective commonsense psychology to make sense of the behavior of another individual what we are doing is surely much closer to the empathetic and hermeneutic application of general tactics, patterns and strategies to a chess player than it is to the identification of the rule-governed framework within which those strategies are applied.

There is a more general issue here, to do with the very possibility of working backwards from what people say they are doing to what they really

are doing. Why should one think that subjects are reliable guides to the principles they are implicitly employing? It is helpful to think about the analogy with other domains where it seems plausible that we employ commonsense or folk theories. Take intuitive physics, for example. We all have from early infancy onwards a set of practical skills that allow us to predict, explain and manipulate the behavior of physical objects – to move around without bumping into objects whether those objects are at rest or in motion, to calculate on the basis of very incomplete information the trajectory and speed of moving objects in a way that allows us to avoid, intercept or follow them. These are physical counterparts of the practical skills and abilities of unreflective commonsense psychology. It is just as plausible in the case of intuitive physics as it is in the case of commonsense psychology to think that these practical skills and abilities are underwritten by a set of implicit beliefs about the behavior of moving and stationary objects. The nature of these implicit beliefs has been systematically studied by experimental psychologists (e.g. McCloskey 1983). One striking feature of this research has been the dissociation it has revealed between the expectations that people have and the principles to which they verbally assent – between what they actually do and the principles that they are "able and willing to acknowledge as correct", to use Dummett's phrase. When we look at what people say about the behavior of objects we find a striking number of basic misconceptions and errors (McCloskey 1983). These misconceptions and errors at the reflective level do not carry over, however, to the unreflective way in which people interact with the physical world. People would have real problems if they actually behaved in accordance with the principles that they explicitly accept.

In one well-known experiment from the literature on intuitive physics subjects were asked to predict the trajectory that an object would take after exiting a C-shaped tube lying flat on a table. The correct answer is that the object will exit the tube in a straight line, following a trajectory determined by the tangent of the tube's curvature at the point of exit. As is often the case in these experiments the subjects were college students, whom one might expect to be reasonably educated and sophisticated. Of the students studied by McCloskey *et al.* (1980), only 60 percent made the correct prediction, with 40 percent predicting that the object would follow a curving trajectory that continued the curve of the tube. The proportion of incorrect answers is striking. It seems unlikely, however, that it is associated with widespread practical difficulties. One would not, for example, expect 40 percent of the population to make the corresponding error when it came to catching a ball exiting a C-shaped tube. Some confirmation can be found in experiments that have used animation to present subjects with contrary-to-fact states of affairs corresponding to the explicit predictions that they made about object motion (Kaiser *et al.* 1986). Subjects tend to describe situations in which, for example, objects follow curvilinear trajectories as looking odd.

The example of intuitive physics suggests a degree of skepticism about

using the principles and judgments to which people explicitly assent as a reliable guide to what is going on in their unreflective practice. It looks as if we should be wary of taking verbal reports and commonly accepted principles and platitudes to be a guide to the content of unreflective commonsense psychology. In view of this, and of the general unclarity of the notion of implicit knowledge (as applied to commonsense psychology), it will be helpful to look at alternative ways of understanding the scope of commonsense psychology. The issue here is just as much one of empirical matter of fact as it is of conceptual analysis or philosophical argument. The question is how social understanding and social coordination actually work. What are the psychological mechanisms that underlie social coordination and social understanding? Of course, the question is not purely empirical, because we still need a proper theoretical articulation of the different possibilities. Nonetheless, the philosophical and analytical issues are much more closely tied to empirical issues than philosophers have generally been prepared to admit.

The remainder of this chapter explores alternatives to propositional attitude psychology, as it is standardly understood. Section 7.3 considers a way of thinking about propositional attitude psychology that tries to avoid attributions of implicit knowledge. This is the *simulationist* approach to propositional attitude psychology, which takes social understanding and social coordination to rest upon capacities to simulate the cognitive and emotional perspective of others – to think about what one would do if one were in their position. In an important sense, however, simulationism remains a version of propositional attitude psychology. In section 7.4 we look at ways in which social understanding and social coordination might proceed in complete independence of propositional attitude psychology.

7.3 Modest revisionism: the simulationist proposal

We have been looking in this chapter at a particular type of proposals about how commonsense psychology can be applied in everyday social situations. These are proposals that appeal to the notion of *implicit knowledge*. Theorists as disparate as Lewis, Braddon-Mitchell and Jackson, Fodor and Paul Churchland accept the view that social understanding and social coordination rest upon an implicitly known, and essentially theory-like, body of generalizations connecting propositional attitude states with overt behavior and with each other. Social understanding involves subsuming observed behavior and what is known of a person's mental states under these generalizations in order to understand why they are behaving in a certain way and how they will behave in the future. This picture of how propositional attitude psychology works has come to be known as the *theory-theory* (that is, the theory that propositional attitude psychology takes the form of a theory). Theorists promoting the simulationist approach to commonsense psychology have challenged the theory-theory in recent years within both philosophy and psychology (Gordon 1986; Heal 1986).

Simulationists think that we explain and predict the behavior of other agents by projecting ourselves into the situation of the person whose behavior is to be explained/predicted and then using our own mind as a model of theirs. Suppose that we have a reasonable sense of the beliefs and desires that it would be appropriate to attribute to someone else in a particular situation, so that we understand both how they view the situation and what they want to achieve in it. And suppose that we want to find out how they will behave. Instead of using generalizations about how mental states typically feed into behavior to predict how that person will behave, the simulationist thinks that we use our own decision-making processes to run a simulation of what would happen if we ourselves had those beliefs and desires. We do this by running our decision-making processes *off-line*, so that instead of generating an action directly they generate a description of an action or an intention to act in a certain way. We then use this description to predict the behavior of the person in question.

The simulation theory is most often presented in the context of prediction, and it is less clear how it works in the service of explanation. Prediction works in the same direction as our ordinary decision-making processes – both prediction and decision-making involve processes of transforming mental states into behavior. Explanation, on the other hand, works in the opposite direction to decision-making. What we are trying to do in psychological explanation is work backwards from behavior to the causes of behavior. The simulationist idea, presumably, is that we run our decision-making processes off-line, using a range of different pairs of beliefs and desires, until we come up with a belief–desire pair that produces something close to the observed behavior. We then infer that that belief–desire pair produced the behavior in question.

The issue separating the theory-theorist and the simulationist is not primarily the scope of commonsense psychology, although as we shall see the simulation theory does have implications for this question. Rather, the important issues are (a) how we arrive at the attributions of beliefs and desires, and (b) how we get from those attributions to explanations/predictions. The following passage from Gregory Currie makes clear both the differences between the two positions and the ground they share.

> Simulation theorists say that our access to the thoughts of others is not through the application of a primitive but effective theory, as advocates of the "theory-theory" of folk psychology suppose, but through a kind of internal, largely spontaneous, re-enactment that allows us to imagine ourselves in some rough approximation to the situation of another. In so imagining, we tend to acquire, in imagination, the beliefs and desires an agent would most likely have in that situation, and those imaginary beliefs and desires have consequences in the shape of further pretend beliefs and desires as well as pretend decisions that mimic the beliefs, desires and decisions that follow in the real case.
>
> (1995 p. 158)

Both theory-theorists and simulation theorists, therefore, think that we tend to arrive at predictions and explanations by moving from beliefs and desires, either through theoretical principles that link particular complexes of beliefs and desires to particular behaviors or through working out what one would oneself do in that situation with those beliefs and desires.

We should distinguish two ways of developing this basic simulationist idea. On one version, the process of simulation still requires the *explicit* attribution of beliefs and desires to the person being simulated. On this view, in order to simulate someone I need to form explicit judgments about how they represent the relevant situation and what they want to achieve in that situation. These judgments serve as the input to the simulation process. I might reach these judgments by thinking about the beliefs and desires I myself would have in a particular situation. I might exploit my knowledge of a particular person's "take" upon the world. I might even use some rule of thumb to make a hypothesis about how the world might look from that person's vantage point. But, whatever mechanism I employ, I will nonetheless have to make an explicit attribution of propositional attitude states to the person in question. This version of simulationism clearly involves deploying the conceptual framework of propositional attitude psychology – and it should be clear (as we will see further below) that it opens the door for the theory-theorist to object that this process rests upon implicit knowledge of psychological generalizations. I will call it *standard simulationism*.

But there is room for a second way of developing the basic simulationist idea – one that does not assume that we need to deploy the conceptual framework of commonsense psychology in order to arrive at inputs to the process of simulation. The intuitive idea here is that, instead of coming explicitly to the view that the person whose behavior I am trying to predict has a certain belief (say, the belief that *p*), what I need to do is to imagine how the world would appear from his point of view (Gordon 1986). Let us call this *radical simulationism* – as opposed to standard simulationism. The distinction is between, on the one hand, forming a belief about how another person represents the world (a belief with the content that the person believes that *p*) and, on the other, holding a belief about the world in one's imagination (a belief with the content simply that *p*). One central point at issue is whether a simulation involves deploying the concept of belief and thinking about the beliefs that another person might have. According to standard simulationism, I cannot simulate the beliefs of another without possessing the concept of belief, and my simulation is directed primarily at the other person's psychological states. According to radical simulationism, on the other hand, what the simulator is thinking about is the world, rather than the person they are simulating. The simulator is thinking about the world *from the perspective of the person being simulated*, rather than thinking about their beliefs, desires and other psychological states. The spirit of this "world-directed" way of thinking about psychological explanation comes across in the following passage from Jane Heal, one

of the leading simulation theorists (although she prefers to talk about *replication*, rather than *simulation*).

> On the replicating view psychological understanding works like this. I can think about the world. I do so in the interests of taking my own decisions and forming my own opinions. The future is complex and unclear. In order to deal with it I need to, and can, envisage possible but perhaps non-actual states of affairs. I can imagine how my tastes, aims, and opinions might change, and work out what would be sensible to do or believe in the circumstances. My ability to do these things makes possible a certain sort of understanding of other people. I can harness all my complex theoretical knowledge about the world and my ability to imagine to yield an insight into other people *without any further elaborate theorizing about them*. Only one simple assumption is needed: that they are like me in being thinkers, that they possess the same fundamental cognitive capacities and propensities as I do.
>
> (Heal 1986, reprinted in Davies and Stone 1995a, p. 47)

At the moment, therefore, we have three different theories in play of how psychological explanation proceeds – the theory-theory, standard simulationism and radical simulationism. They each offer different ways of understanding unreflective commonsense psychology at the third of the three levels identified earlier. At the top level, unreflective commonsense psychology is simply the complex of skills and abilities (whatever they might turn out to be) that underlie social understanding and social coordination. At the second level, we can understand unreflective commonsense psychology in slightly more determinate terms, namely, as involving the conceptual framework of propositional attitude psychology. At third level we have different ways of understanding how that conceptual framework is actually applied, as shown in Figure 7.1.

We can get a firmer sense of how the three different theories about the practical application of commonsense psychology might be applied in practice by looking at the different interpretations they provide of one of the key

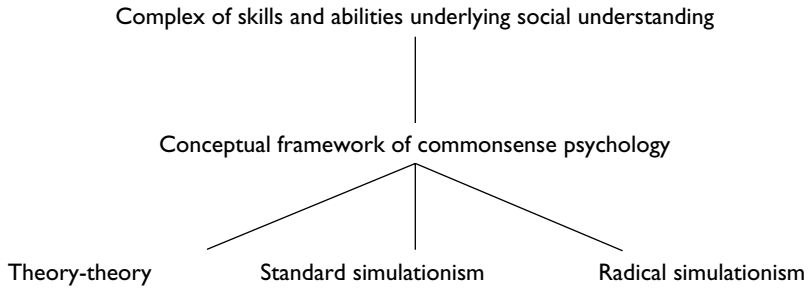

Figure 7.1 Three ways of thinking about commonsense psychology.

psychological experiments in the psychological literature on the development of psychological understanding in young children – the false belief task, as developed by Wimmer and Perner (1983).

The false belief task explores young children's understanding of the possibility that someone might have mistaken beliefs about the world. What is distinctive about the mental state of belief (as opposed, for example, to perception) is that beliefs can be true or false. There is no sense in which one can perceive that something is the case without it actually being the case – any more than one can know that *p* without it being the case that *p*. Both knowledge and perception *track* the world – they are what philosophers sometimes call *factive states*. In contrast, the way beliefs represent the world has no such implications for how the world actually is. With this distinction in mind, it seems very plausible that there is no sense in which someone can understand what belief is (can possess the concept of belief) without understanding that having a belief is representing the world in a way that could turn out to be false.

The false belief task is intended to identify whether a child properly grasps this crucial dimension of the concept of belief. The task has by now appeared in many different forms, but in its original formulation it was based on a puppet show featuring a child called Maxi and his mother. Young children are shown a short puppet show in which Maxi hides some chocolate in a box and then goes out to play. While he is out his mother moves the chocolate from the box to the cupboard. The question put to the children is: where will Maxi look for the chocolate when he gets back? The choice is between the box (where Maxi should still believe the chocolate to be, since he is completely unaware that his mother has moved it) and the cupboard (where the chocolate actually is, and where the child watching the show knows it to be). It turns out (and the data are quite robust) that up to the age of about four children answer that Maxi will look in the cupboard. Between four and five, however, most non-autistic children arrive at the correct answer.

The three theories that we have been considering will interpret the transition noted by the false belief task in different ways. The theory-theory, which is the dominant view among developmental psychologists studying the false belief and related tasks, maintains that children who fail the false belief task have not yet acquired the concept of belief. On this view, the concept of belief is defined by its role in certain generalizations. One such generalization might be that people's beliefs change only when they receive new information, so that information of which they are unaware will have no impact on their beliefs. Failure on the false belief task shows that the young child's embryonic psychological theory does not yet encompass the concept of belief. The way in which failure on the false belief task would be interpreted by standard simulationism is somewhat similar. Like the theory-theory standard simulationists would stress the child's inability to represent Maxi's beliefs. Children who fail the false belief task do not have the appropriate inputs for their simulations. The reason for this is not, however, that

they lack the appropriate theoretical knowledge. Rather, children below the age of four are not yet capable of simulating the process of acquiring a belief – the process of moving from a perceptual state that tracks the world to a perceptual belief that does not track the world. In the case of radical simulationism, however, the problem is not at all to do with the concept of belief. Radical simulation is world-directed, rather than mind-directed. For the radical simulationist the problem with children who have failed the false belief task is that their capacity to project themselves imaginatively is not yet sufficiently developed. They are not yet able to form beliefs from a point of view other than their own. They are capable of imaginatively perceiving (of taking Maxi's perceptual perspective on the world), but not of imaginatively taking Maxi's *doxastic* perspective on the world.

There is an extensive empirical debate about the respective advantages of the theory-theory and the simulation theory. The assumption underlying the debate is that the theory-theory and the simulation theory generate testably different predictions. Consider, for example, the following argument put forward by Stephen Stich and Shaun Nichols (Stich and Nichols 1995). According to the simulation theory, the process of predicting how someone will behave in a particular situation essentially uses the same mechanisms as the process of making up one's own mind about how to act in that very same situation. Consider, therefore, a situation in which the overwhelming majority of people react in a similar way and yet where their behavior seems puzzling and perhaps even irrational. Take, for example, the position effect in the selection of consumer goods. It turns out that when subjects are asked to rate the quality of an array of goods that are indistinguishable from each (but that they do not know to be indistinguishable from each other) there is a very robust tendency to opt for items on the right-hand side of the display.[5] One might intuitively think that the sensible thing to do would be to make a random choice, and no doubt that is what the subjects think that they are doing – as it turns out, however, their putatively random selections are not so random after all. Suppose, now, that we consider, not what people actually do in situations such as those, but rather what they will *predict* that other people will do in those situations. Proponents of the theory-theory, such as Stich and Nichols, think that predictions will be based upon some sort of principle similar to that just discussed – namely, that selections will be random. So, the predictions will be largely mistaken, predicting a fairly random distribution when in fact there is a heavy concentration towards the right-hand end of the array. But what about the simulation theory? Stich and Nichols think that, since the position effect is very common, the predictor would be likely to make the same choice as the people whose behavior she is trying to predict. Suppose, then, that she runs her own decision-making processes off-line in order to simulate the behavior of the people choosing between, say, identical washing machines. Stich and Nichols

5 This puzzling phenomenon, along with many others, is discussed in Nisbett and Ross (1980).

suggest that she will make a correct prediction – because her prediction will track her own dispositions to behave. It turns out that subjects tend to be very bad at predicting things like the position effect, which Stich and Nichols take as an argument for the theory-theory.

It is unlikely that arguments such as these will ever be conclusive. The predictor may well not be in a comparable position to the person whose behavior he is trying to predict. The predictor may, for example, have information that the person making the choice does not have – such as the knowledge that there are no differences whatsoever between the items on display. This information might feed into the simulation so that the predictor's prediction no longer tracks how she herself would have behaved in that situation. Nor of course is the simulator ever in exactly the same situation as the person whose behavior he is predicting. He is not confronted with the array of indistinguishable washing machines, but merely has them described to him. Of course, the theory-theorist encountering these objections is likely to reply that they save the battle only at the cost of losing the war. If a predictor needs to have exactly the same information, and in the same format, as the person they are predicting, then we can expect simulation to be a relatively infrequent occurrence.

Putting empirical debates to one side, however, the simulation theory does promise some definite advantages over the theory-theory. One key problem for theory-theorists is that the generalizations and rules of thumb that they think govern everyday social interactions and commonsense psychological explanations hold only for the most part. All commonsense psychological generalizations have exceptions and even when they do apply to a given situation they do so only in a *prima facie* manner. It is perfectly possible for them to be trumped by different generalizations. It might be reasonable to assume, for example, that people will generally do what they think will best further the satisfaction of their desires – but there are all sorts of reasons why someone might not act in that way in a given situation. The generalizations of commonsense psychology hold at best *ceteris paribus* (all other things being equal). This is frequently thought to be problematic for rule-based approaches to commonsense psychological explanation. How are subjects to know whether all other things actually are equal? How are they to work out whether a putative generalization really does apply in a given situation? It seems highly implausible to think that subjects are in some sense aware (even implicitly aware) of all the possible exceptions to a given generalization. In what sense, therefore, should they properly be described as knowing the relevant generalizations?

This line of objection hardly presents an insuperable difficulty for proponents of the theory-theory, who can argue, for example, either that all laws (including scientific laws) are *ceteris paribus* laws or that one can know and apply a *ceteris paribus* generalization without being able to spell out all the possible exceptions. Nonetheless, the simulation theory provides a way of avoiding all such difficulties. It is open to a simulation theorist to argue that

our commonsense psychological generalizations inherit the precision and determinacy of our ordinary decision-making processes. Whatever mechanisms secure disambiguation and resolution of conflicts between different reasons for acting in our ordinary processes of decision-making will equally secure disambiguation and resolution of conflicts when we are trying to explain the behavior of others. There is a sense in which this postpones the problem rather than resolving it – after all, we know very little about how our decision-making processes actually work. But nonetheless, we do know that the problem has a solution, since we do know that our decision-making processes tend to produce unique solutions to problems. And we also know that our decision-making processes might operate without any explicitly coded generalizations – and consequently without there being any need to explain how the *ceteris paribus* clauses are known.

There is a range of theoretical objections to the simulation theory. Perhaps the most significant concerns the extent to which the simulation theory is really different from the theory-theory. We have already seen that the simulation theory is able to trade on our *de facto* ignorance of the details of how our actual decision-making processes work. The key point for the simulation theory is that those processes, *whatever they turn out to be*, work both to generate our own actions and to explain/predict the actions of others. But the simulation theorist has to leave open, of course, the possibility that those processes might turn out to involve some form of tacitly known psychological theory. Suppose, for example, that our decision-making processes essentially involve calculating expected utility. Calculations of expected utility take place within the theoretical framework of decision theory (or some psychologically plausible version thereof). Suppose, then, that, as the simulationist suggests, we explain/predict the behavior of others by running our own decision-making processes off-line. We would, therefore, be using expected utility theory to predict other people's behavior. How exactly does this account of psychological explanation differ from that, for example, of a theory-theorist such as David Lewis, who thinks that decision theory is a regimentation of tacitly known commonsense psychology? Exactly the same mechanisms seem to be in play in both cases.

The only possible difference between the simulation theory and the theory-theory on this scenario would be in how the explainer/predictor arrives at the appropriate inputs for the calculation of expected utility. The theory-theorist would say that we use various rules of thumb and psychological generalizations to work out people's utility and probability assignments on the basis of what they do and say. Does the simulation theorist have a competing account? It seems plausible that the radical simulationist does indeed have a competing account. Probability and utility assignments for the other person are derived by imaginative adopting of their point of view. These assignments are then fed directly into the calculations of expected utility. But it is not clear that the standard simulationist has so clear an alternative to offer. The standard simulationist is committed to holding that

we form beliefs about the other person's utility and probability assignments. But where do these beliefs come from? A theory-theorist would argue that the standard simulationist will have to appeal to precisely the same psychological generalizations and rules of thumb that someone like Lewis would use to arrive at utility and probability assignments.

Nor is the radical simulationist entirely immune to the objection that his position is in danger of collapsing into the theory-theory. Radical simulationists have to explain how we think ourselves into another person's point of view – how we adopt their perspective on the world. Clearly, the more similar they are to us the easier this process will be. But the difficulty comes when we think about how one might compensate for differences. In some cases, when we are dealing with people whom we know well, it is easy enough to make the necessary adjustments that will allow us to adopt their perspective on the world – we can pretend to be more altruistic than we really are, for example, in order to think our way into the beliefs and desires of someone whom we happen to know is very selfless. But what happens when we are dealing with people whom we do not know at all? How are we to make the relevant adjustments? How are we to modify our own perspective on the world in order to be able to think our way into theirs? Once again it is easy to see how the theory-theorist is likely to argue, namely, that making these adjustments is only possible if we bring to bear a body of theoretical generalizations about psychological states, how they emerge in response to different situations and how they interact with each other to generate behavior.

It is in many ways puzzling that the debate between the theory-theory and the simulation theory has taken such a stark form. One might wonder why one has to make a choice between one of the two approaches. Is there really likely to be one single account of how we employ propositional attitude psychology to explain and predict other people's behavior? Perhaps we should be thinking instead of a spectrum of possible modes of application. Some of these might fall closer to a "pure" theory-theory and some to a "pure" simulation theory, but it might well be the case that much of the time when we employ propositional attitude psychology we employ a combination of simulation and theory (Heal 1996; Perner 1996). We might empathetically think our way into someone's point of view to try to understand their beliefs and desires and then use theoretical generalizations to derive an explanation or prediction. Alternatively, we might deploy our theoretical knowledge to understand someone else's perspective and then use what we would ourselves have done had we had that perspective to move to an explanation or prediction.

Returning to the scope of commonsense psychological explanation, only radical simulationism offers a way of developing the thought that our unreflective social understanding and social coordination might not rest exclusively upon deploying the concepts and categories of propositional attitude psychology. The whole point of radical simulationism is that we can explain,

predict and interact with other people without thinking about their beliefs and desires – we merely think ourselves into their position. So, radical simulationism gives us one way of thinking about how the domain of commonsense psychology might actually be narrower than it is standardly taken to be. Are there any other ways in which social understanding and social coordination might take place without involving the machinery of propositional attitude psychology? We explore this possibility in the next two sections.

7.4 Narrowing the scope of commonsense psychology (1)

There are some very general reasons for thinking that commonsense psychology cannot be as dominant as it is taken to be. Some stem from considerations of cognitive architecture and the structure of the mind. Others stem from considerations of computational complexity. We will look at both in this section, moving on in the next section to consider some practical alternatives to commonsense psychology.

The computational argument is straightforward. It is motivated by the thought that the vast majority of our social interactions involve almost instantaneous adjustments to the behavior of others, whereas folk psychological explanation is a complicated and protracted business, whether it is understood according to the simulation theory or the theory-theory. It is no easy matter to attribute beliefs and desires and then to work either backwards from those beliefs and desires to an explanation or forwards to a prediction. The point is easiest to see with respect to the theory-theory. To apply folk psychological explanation is to subsume observable behavior and utterances under general principles linking observable behavior to mental states, mental states to other mental states and mental states to behavior. As many authors have stressed, we can only apply these principles if we can identify, among a range of possible principles that might apply, the ones that are the most salient in a given situation. We need to identify whether the appropriate background conditions hold, or whether there are countervailing factors in play. We need to think through the implications of the principles one does choose to apply in order to extrapolate their explanatory/predictive consequences. The need to do all these things makes folk psychological generalizations rather unwieldy. And it is no surprise that the paradigms of folk psychological explanations given by theory-theorists tend to be complicated inferences of the sort either found in the final chapters of detective novels (e.g. Lewis 1972) or in dramatic and self-questioning soliloquies (e.g. Fodor 1987, Chapter 1). These are striking cognitive achievements, but it seems odd to take them as paradigms of interpersonal cognition. Do our everyday cognitive interactions with people really involve deducing hypotheses from general principles, drawing out the deductive consequences (more accurately: the *relevant* deductive consequences) of those

general principles and then putting those hypotheses before the tribunal of experience? If that is what is required then it is a wonder that such a thing as social coordination exists.

The narrow range of examples that tend to be considered may well obscure the practical difficulties here. Folk psychological explanation is usually considered by philosophers to be a one-on-one activity. This is exactly what one would expect given that the paradigms are the detective drawing together the strands of the case, or the puzzled lover trying to decode the behavior of her paramour. But social understanding is rarely as circumscribed as this. In many examples of social coordination there is a range of people involved and the behavior of any one of them is inextricably linked with the behavior of the others. Suppose that the social understanding involved in such examples of social coordination is modeled in commonsense psychological terms. This would require each participant to make predictions about the likely behavior of other participants, based on an assessment of what those participants want to achieve and what they believe about their environment. For each participant, of course, the most relevant part of the environment will be the other participants. So, my prediction of what another participant will do depends upon my beliefs about what they believe the other participants will do. The other participant's beliefs about what the other participants will do are in turn dependent upon what they believe the other participants believe. And so on.

There will be many layers in the ensuing regress, and the process of coming to a stable set of beliefs that will allow one to participate effectively in the coordinated activity will be lengthy and computationally demanding. Of course, none of this shows that there are any objections in principle to modeling coordinative social understanding in folk psychological terms. Any such claim would be absurd, not least because we have a well worked out mathematical theory that allows us to model social understanding in what are essentially folk psychological terms (or at least a regimentation of them). Game theory is a theory of social coordination and strategic interaction employing analogs of the folk psychological notions of belief and desires (in the guise of probability and utility assignments). What thinking about computational tractability should do, however, is at least to cast doubt upon whether this could be a correct account of the form of social understanding in the vast majority of situations.

It is important to distinguish this point from another charge leveled at the theory-theory. Simulation theorists have sometimes suggested that issues of computational tractability work in favor of the simulation theory. Jane Heal, for example, has argued that theory-theorists run into difficulties analogous to the frame problem in computer science (Heal 1996). The frame problem is essentially the problem of determining which, among the myriad aspects and deductive consequences of a principle or of a belief, are relevant in a given situation (Dennett 1984 and pp. 26–27). Any psychological theory incorporating a satisfactory response to the frame problem will of

necessity incorporate a theory of relevance, specifying why certain psychological factors will be deemed relevant in some situations but not in others, how changing the parameters of a situation can radically alter those aspects of it relevant to decision-making; and how what is taken to be relevant can vary systematically with determinate aspects of the psychology of the individual. It is, according to Heal, a weighty consideration against the theory-theory that any such, presumably tacitly known, theory of relevance would be far more complex than any other postulated tacit theory to explain, for example, our grasp of grammar or of so-called naïve physics.

This worry is well grounded (although one might wonder whether a simulation theorist can avoid postulating at some level a tacitly known theory of relevance governing both our on-line decision-making processes and our off-line simulations). But it is orthogonal to the computational worry we are considering. That computational worry would still be there even if we granted the theory-theorist the legitimacy of postulating a tacitly known theory of relevance. The worry about relevance is a worry about how it is even possible to tailor the generality of folk psychological principles to the particularity of specific situations. The computational worry, on the other hand, is about the combinatorial explosion that will occur when the situation in question involves several individuals who are potentially collaborating. Even if we can fix the parameters of relevance in a way that will permit folk psychological principles to come into play, the key problem comes from the fact that the application of folk psychological explanation to a multi-agent interaction will require a computationally intractable set of multiply embedded higher-order beliefs about beliefs.

The worry about combinatorial explosion is not confined to the theory-theory. Let us suppose that the simulation theory can get by without having to assume a tacitly known theory of relevance, so that a simulation simply involves using one's own mind as a model of the minds of the other participants in the interaction. One would still need to plug into the decision-making processes an appropriate set of inputs for all the other participants and then run simultaneous simulations for all of them. This is multiply problematic. There is, first of all, a straightforward question about how many simulations it is actually possible to run simultaneously. Since the practical details of how the process of simulation might work have not really been explored, there is little concrete to say about this. *Prima facie*, however, one might think that there will be some difficulties with the idea of multiple simultaneous simulations, given that a simulation is supposed to work by running one's own decision-making processes off-line and those processes are presumably designed to give an output for a single set of inputs. But there is a more serious problem. The simultaneous simulations will not be independent of each other. Suppose that the interaction contains three participants, *A*, *B* and *C*, in addition to me. In order to simulate *B* properly I will need to have views about what *A* and *C* will do – without that information I will not have any sense of what initial beliefs it would be

reasonable to attribute to *B*. But, by parity of reasoning, this information about what *A* and *C* will do depends upon each of them having information about what the other participants will do. It is very difficult to see how the notion of simulation can be stretched to accommodate, not just simultaneous simulations, but simultaneous simulations that are interdependent. So, the simulation theory, no less than the theory-theory, is bound to confront problems of computational tractability if it adopts a broad construal of the domain of commonsense psychology.

Let us turn now to a second general reason for skepticism about the broad interpretation of the domain of commonsense psychology. Here I will be painting with very broad strokes of the brush indeed. Folk psychological reasoning is a paradigm of metarepresentational thinking, where metarepresentational thinking involves thinking about thoughts – taking thoughts as the objects of thought, attributing them to other subjects, evaluating their inferential connections with other thoughts, and so on. It has been suggested that metarepresentational thinking is in some sense language-dependent (Dennett 1996; Bermúdez 2003a, Chapter 8, and see Chapter 10 in this volume for further discussion). One might argue, for example, that thoughts must have vehicles that are consciously and reflectively accessible if they are to feature in metarepresentational thinking, and that the only possible vehicles are linguistic.[6] If the thesis of language-dependence is correct, then it seems likely, on the basis of our best current theories of cognitive archeology, that many of the cognitive skills involved in social coordination emerged long before the capacity for metarepresentational thinking, and hence long before folk psychological explanation was even possible.[7] Early hominids, whom we do not believe to have possessed language, appear to have been capable of an impressive range of types of collective behavior, involving the social transmission of knowledge (e.g. knowledge of the natural world); the tracking of social relations within social groups; complex forms of social coordination (in hunting and migratory behavior) and technical training in tool manufacture (Mithen 1996). All these forms of social coordination require high degrees of social understanding. *Ex hypothesis* this social understanding could not have involved the concepts and explanatory/predictive strategies of commonsense psychology.

Of course, this does not allow us to draw any immediate inferences about the current state of our social cognition – perhaps the metarepresentational abilities that emerged with language acquisition (or at any rate relatively late in cognitive evolution) simply wrote over their primitive precursors, in the way that some developmental psychologists think that the earliest conceptions of the physical world acquired in infancy are completely superceded by the "naïve physics" emerging later in development (Gopnik and Meltzoff

6 See section 10.2.

7 In fact, this suggestion is independently plausible even without the thesis of language-dependence.

1997). That would doubtless be the position of those who adopt what I have termed the broad construal of the domain of commonsense psychology. But much of what we know about the evolution of cognition suggests that this may not have happened. Evolution works by tinkering, grafting new structures onto already existing ones, changing the function of structures that are already there. There is considerable evidence that our cognitive architecture is a patchwork of superimposed structures of varying phylogenetic pedigree.

The points about computational tractability made earlier in this section offer further reasons for thinking that these primitive structures have not only persisted but in fact continue to play an important role in our social lives. It may well not be feasible to think that all or even most of our social interactions can be modeled in commonsense psychological terms. Much of our current social cognition may reflect a residue of skills and abilities that long preceded the emergence of metarepresentation and commonsense psychology. There is little to be gained, however, from pursuing this line of thought without providing concrete examples of the form that these skills and abilities might take. We will turn to that task in the next section.

7.5 Narrowing the scope of commonsense psychology (2)

This section considers three different examples of how social understanding and social coordination might be secured without recourse to the conceptual framework of commonsense psychology.

The first example shows how social interactions can depend upon participants' sensitivity to each other's emotional states without those participants explicitly attributing emotional states to one another. Navigating the social world is often a matter of being directly sensitive to the emotional states of others without making any explicit judgments about those emotional states (and hence, *a fortiori*, without either simulating them or theorizing about them). The second shows (in an idealized way) how social interactions might proceed without the participants having either to explain or to predict the behavior of other participants. It may well be that a significant amount of social behavior is governed by simple algorithms that allow speedy decision-making without the complexities of predicting how other people have behaved, or explaining why they behaved they way we did. We will look at the use of the TIT-FOR-TAT algorithm in thinking about the prisoner's dilemma as an example of how this might work. Third, we explore social routines and frames as examples of how subjects might make predictions about the behavior of other subjects (and indeed offer retrospective explanations for their actions) without attributing psychological states or bringing to bear any of the machinery of commonsense psychology.

Emotion perception in social interactions

It has been known for some time that emotion perception is highly dependent upon cues operating far below the threshold of conscious awareness. Emotional states can be transmitted directly from person to person. This plays an important role in many types of social interaction, particularly those involving collective behavior. We have a reasonably worked out understanding of how this transmission of emotional states can take place. The role of facial expression in the communication and detection of emotion has been systematically studied since Charles Darwin's pioneering study (Darwin 1872). Recent neuroscientific research based on the study of brain-damaged patients and on lesion studies in animals has postulated the existence of neural circuits dedicated to the production and understanding of expressive behavior, the so-called limbic system (Ledoux 1996).

The simple claim that emotion perception is frequently subliminal does not count against the broad conception of commonsense psychology. Directly perceived emotional states can easily serve as inputs to the processes of simulation, or as the raw material to which the generalizations of theoretical folk psychology are applied. The more interesting, and controversial, suggestion is that we frequently act upon the perception of emotional and affective states without explicitly identifying them. The idea here is that we regulate our own behavior as a function of our sensitivity to the emotional and affective states of those with whom we are interacting without at any point making explicit the identifications on which our behavior rests. The understanding of emotional expression feeds directly into behavior. Sensitivity to emotional states feeds directly into action without any attribution of emotional states.

But even in this sort of social situation, the issue is often not what other participants will do but how they will do it. Situations where emotion perception is important are rarely situations where issues of explanation and prediction arise in the sort of ways that seem to require folk psychological forms of social understanding. In any case, the fact that many social interactions involve an element of "affect attunement" (Stern 1985) achievable without recourse to folk psychology hardly shows that no element of those interactions is controlled folk psychologically. What we need to ask now is whether there are interpersonal situations that are *not* circumscribed by shared goals or a relatively small number of clearly defined possible outcomes and yet where we can act effectively *without* actively explaining and/or predicting the behavior of other participants in terms of what they believe and desire. This brings us to the second example.

The indefinitely iterated prisoner's dilemma

A prisoner's dilemma is any strategic interaction where the dominant strategy for each player leads inevitably to an outcome where each player is worse

off than he could otherwise have been. A dominant strategy is one that is more advantageous than the other possible strategies, irrespective of what the other players do. In the standard example from which the problem derives its name, the two players are prisoners being separately interrogated by a police chief who is convinced of their guilt, but as yet lacks conclusive evidence. He proposes to each of them that they betray the other, and explains the possible consequences. If each prisoner betrays the other then they will both end up with a sentence of five years in prison. If neither betrays the other, then they will each be convicted of a lesser offence and both end up with a sentence of two years in prison. If either prisoner betrays the other without himself being betrayed, however, then he will go free while the other receives ten years in prison. The dominant strategy for each player is to betray the other. Since we are dealing with rational players it follows that each will implicate the other, resulting in both spending five years in prison – even though had they both kept quiet they would have ended up with just two years apiece. We can see how this works by looking at the pay-off table.

The table illustrates the pay-offs for the different possible outcomes of a one-shot prisoner's dilemma. Each entry represents the outcome of a different combination of strategies on the part of prisoners A and B. The bottom left-hand entry represents the outcome if prisoner A keeps silent at the same time as being betrayed by prisoner B. The outcomes are given in terms of the number of years in prison that will ensue for prisoners A and B respectively. So, the outcome in the bottom left-hand box is ten years for prisoner A and none for prisoner B.

		Player B	
		BETRAY	KEEP SILENT
	BETRAY	5, 5	0, 10
Player A			
	KEEP SILENT	10, 0	2, 2

Although some authors have tried to argue otherwise (e.g. Gauthier 1986), it is hard to see how it can be anything but rational to follow the dominant strategy in a *one-off* strategic interaction obeying the logic of the prisoner's dilemma. Imagine looking at the pay-off table from prisoner A's point of view. You might reason as follows.

> Prisoner B can do one of two things – betray me or keep quiet. Suppose he betrays me. Then I have a choice between five years in prison if I also betray him – or ten years if I keep silent. So, my best strategy if he betrays me is to betray him. But what if he keeps silent? Then I have got a choice between two years if I keep quiet as well – or going free if I betray him. So, my best strategy if he keeps quiet is to betray him. Whatever he does, therefore, I'm better off betraying him.

Unfortunately, prisoner B is no less rational than you are and things look exactly the same from her point of view. In each case the *dominant* strategy is to defect. So, you and prisoner B will end up betraying each other and spending five years each in prison, even though you both would have been better off keeping silent and spending two years each in prison.

Things get more complicated when we come to social interactions that have the same logic as the prisoner's dilemma but are repeated. This creates the possibility of one player rewarding another for not having betrayed him. One might think that this will change what it is rational to do. But it only does so in a limited range of situations. The so-called backwards induction argument suggests that the rational course of action where each player is rational, knows the other player to be rational and is certain in advance how many strategic interactions there will be is to defect on the first play.[8] But when it is not known how many plays there will be and/or the rationality of the other participant is not known, that scope opens up for cooperative play.

This is where we rejoin the question of the domain of commonsense psychology. Suppose that we find ourselves, as we frequently do, in social situations that have the structure of an indefinitely repeated prisoner's dilemma. The issue may simply be how hard one pulls one's weight in the philosophy department.[9] It may be to my advantage to cut the examination meeting, provided that my colleagues do my work for me. But how will that affect their behavior when we next need to wine and dine a visiting speaker? Will I find myself dining *tête-à-tête* and footing the bill on my own? Before I decide whether or not to cut the examination meeting I had be sensitive to that possibility, and to all the other possibilities when some or all of us applying dominance reasoning will lead to a sub-optimal outcome. But how do I do this?

One answer is that I might make a complex set of predictions about what my colleagues will do, based on my assessment of their preference orderings and their beliefs about the probability of each of us defecting as opposed to cooperating, and then factor in my own beliefs about how what will happen in future depends upon whether or not I come to the examination meeting – and so on. This, of course, would be an application of the general explanatory framework of folk psychology, on the simplification that utilities and probability assignments are regimentations of desires and beliefs – see Pettit

8 The argument is straightforward. Consider the final play. If each player knows that it is the final play, then neither has any reason not to play their dominant strategy. Hence each will betray the other. Consider the penultimate play. Each player is rational and knows the other player to be rational. So each knows what will happen in the final play. This means that they treat the penultimate play as if it were the final play – and hence play their dominant strategy. Exactly the same line of argument holds for the antepenultimate play – and indeed for each play back to the first play. So the outcome on the first play will be mutual betrayal.

9 This is not, strictly speaking, a prisoner's dilemma, since it involves more than two players. The multi-person equivalent of the prisoner's dilemma is usually known as the tragedy of the commons.

(1991), for discussion of the relation between decision theory and commonsense psychology.

But even if we can make sense of the idea that strategic interaction involves these kinds of complicated multi-layered predictions involving expectations about the expectations that other people are expected to have, one might wonder whether there is a simpler way of determining how to behave in that sort of situation. And in fact game theorists have directed considerable attention to the idea that social interactions taking the form of indefinitely repeated prisoner's dilemmas might best be modeled through simple heuristic strategies in which, to put it crudely, one bases one's plays not on how one expects others to behave but rather on how they have behaved in the past. The best known of these heuristic strategies is TIT-FOR-TAT, which is composed of the following two rules:

A. Always cooperate in the first round
B. In any subsequent round do what your opponent did in the previous round

The TIT-FOR-TAT strategy is very simple to apply, and does not involve any complicated folk psychological attributions or explanations/predictions. All that is required is an understanding of the two basic options available to each player, and an ability to recognize which strategy has been applied by other player(s). The very simplicity of the strategy explains why theorists have found it such a potentially powerful explanatory tool in explaining such phenomena as the evolutionary emergence of altruistic behavior (see Axelrod 1984, for an accessible introduction and Maynard Smith 1982, and Skryms 1996, for more detailed discussion).[10]

It is not just that strategies such as TIT-FOR-TAT do not involve any exploitation of the categories of folk psychology. In fact, such strategies do not involve any processes of explanation or prediction at all. In order to apply TIT-FOR-TAT, or some descendant thereof, I need only work out whether the behavior of another player is best characterized as a cooperation or a defection, and which previous behaviors are relevant to the ongoing situation. This will often be achievable without going into the details of

10 TIT-FOR-TAT has only a limited applicability to practical decision-making. In a situation in which two players are each playing TIT-FOR-TAT, a single defection will rule out the possibility of any further cooperation. This is clearly undesirable, particularly given the possibility in any moderately complicated social interaction that what appears to be a defection is not really a defection (suppose, for example, that my colleague misses the examination meeting because her car broke down). So any plausible version of the TIT-FOR-TAT strategy will have to build in some mechanisms for following apparent defections with cooperation, in order both to identify where external factors have influenced the situation and to allow players the possibility of building bridges back towards cooperation even after genuine defection. One possibility would be TIT-FOR-TWO-TATS, which effectively instructs one to cooperate except in the face of two consecutive defections.

why that player behaved as they did. Of course, sometimes it will be necessary to explore issues of motivation before an action can be characterized as a defection or a cooperation – and sometimes it will be very important to do this, given that identifying an action as a defection is no light matter. But much of the time one might get by perfectly well without going deeply at all into why another agent behaved as they did.

Frames and routines

The previous two examples illustrate how one might navigate the social world without explaining or predicting the behavior of others – either through direct sensitivity to other people's emotional states or by using heuristic rules for decision-making. But suppose that neither of these strategies can be used. Suppose that we are dealing with a social situation where some form of explanation and/or prediction of the behavior of other participants is required. Is this a situation that we can only navigate by using the conceptual framework of commonsense psychology? Not necessarily.

Let us start with two very simple examples. Whenever one goes into a shop or a restaurant, for example, it is obvious that the situation can only be effectively negotiated because one has certain beliefs about why people are doing what they are doing and about how they will continue to behave. I cannot effectively order dinner without interpreting the behavior of the person who approaches me with a pad in his hand, or buy some meat for dinner without interpreting the person standing behind the counter. But do I need to attribute folk psychological states to these people in order to interpret them? Must these beliefs about what people are doing involve second-order beliefs about their psychological states? Surely not. Ordering meals in restaurants and buying meat in butcher's shops are such routine situations that one need only identify the person approaching the table as a waiter, or the person standing behind the counter as a butcher. Simply identifying social roles provides enough leverage on the situation to allow one to predict the behavior of other participants and to understand why they are behaving as they are. There is no need to make any folk psychological attributions. There is no need to think about what the waiter might desire or the butcher believe – any more than they need to think about what I believe or desire. The point is not that the routine is cognitively transparent – that it is easy to work out what the other participants are thinking. Rather, it is that we don't need to have any thoughts about what is going on in their minds at all. The social interaction takes care of itself once the social roles have been identified (and I've decided what I want to eat).

One lesson to be drawn from highly stereotypical social interactions such as these is that explanation and prediction *need not* require the attribution of folk psychological states. It would be too strong even to say that identifying

someone as a waiter is identifying him as someone with a typical set of desires and beliefs about how best to achieve those desires. Identifying someone as a waiter is not a matter of understanding them in folk psychological terms at all. It is to understand him as a person who typically behaves in certain ways within a network of social practices that typically unfold in certain ways. This is a case where our understanding of individuals and their behavior is parasitic on our understanding of the social practices in which their behavior takes place. We learn through experience that certain social cues are correlated with certain behavior patterns on the part of others and certain expectations from those same individuals as to how we ourselves should behave. Sometimes we have these correlations pointed out to us explicitly – more often we pick them up by monitoring the reactions of others when we fail to conform properly to the "script" for the situation.

This type of social understanding seems to involve a type of reasoning clearly different from commonsense psychological reasoning as understood by the theory-theory or the simulation theory. For proponents of the theory-theory, social understanding involves what is essentially subsumptive reasoning. Commonsense psychology is a matter of subsuming patterns of behavior under generalizations and deducing the relevant consequences. For proponents of the simulation theory, in contrast, commonsense psychological reasoning is a matter of running one's own decision-making processes off-line and feeding into them appropriate propositional attitude inputs for the person one is interpreting. For those types of social understanding that involve exploiting one's knowledge of social routines and stereotypes, however, the principal modes of reasoning are similarity-based and analogy-based. Social understanding becomes a matter of matching perceived social situations to prototypical social situations and working by analogy from partial similarities. We do not store general principles about how social situations work, but rather have a general template for particular types of situation with parameters that can be adjusted to allow for differences in detail across the members of a particular social category.

Some researchers in computer science defeated by the practical difficulties of trying to provide rule- and logic-based models of commonsense reasoning – difficulties associated with the "frame problem" discussed earlier – have moved towards what are known as *frame-based* systems (Nebel 1999). Here is Minsky's original articulation of the notion of a frame:

> Here is the essence of the theory: when one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a structure called a *frame*. This is a remembered framework to be adapted to fit reality by changing details as necessary.
>
> A *frame* is a data structure for representing a stereotyped situation, like

being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if those expectations are not confirmed.

We can think of a frame as a network of nodes and relations. The top levels of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many *terminals* – slots that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. (The assignments themselves are usually smaller sub-frames.) Simple conditions are specified by *markers* that might require a terminal assignment to be a person, an object of sufficient value, or a pointer to a sub-frame of a certain type. More complex conditions can specify relations among the things assigned to several terminals.

(1974, pp. 111–112)

The frame-based approach is not, of course, confined to the representation of social situations and interpersonal configurations. Frames can have patterns of behavior built into them. They provide a concrete example of the form that a routine-based approach to social understanding and social coordination might take.

We should separate out different possible claims here. Conceding that much of our social understanding may be frame-based rather than rule-based is not automatically to provide a further narrowing of the domain of folk psychology. It may be that the parameters in the frame that need to be set (what Minsky calls the terminals or slots) include specifications of the mental states of the other parties in the interaction. However, it might equally be argued that this will not be the case (or at least will not be the case for many of our frame-based social interactions). The parameters associated with the other participants are set by specifications of roles and behavior, rather than by specifications of beliefs and desires.

7.6 A suggestion?

One conclusion to draw from the examples considered in section 7.5 is that the social world is often transparent, easily comprehensible in terms of frames, social roles and social routines. Other agents can be predicted in terms of their participation in those routines and roles, while their emotional and affective states can simply be read off from their facial expression and the "tenor" of their behavior. When the social world is in this way "ready-to-hand", to borrow from Heidegger's characterization of the practical understanding of tools, we have no use for the apparatus of commonsense psychology. We have no need of it to navigate through the social world, to accommodate ourselves to the needs and requirements of

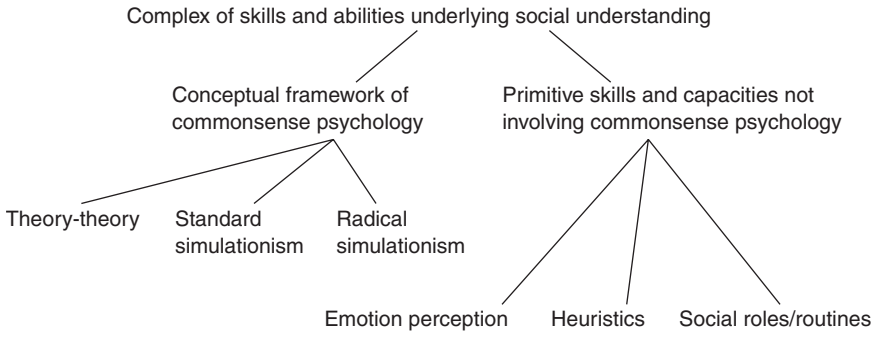

Figure 7.2 Elements of social understanding.

other people and to succeed in coordinated activities. But sometimes the social world becomes opaque. We find ourselves in social interactions where it is not obvious what is going on; that cannot easily be assimilated to prototypical social situations; where we cannot work out what to do simply on the basis of previous interactions with the other participants. And it is at this point, it would be suggested by proponents of the narrow construal, that we find ourselves in need of the type of metarepresentational thinking characteristic of folk psychology – not as a mainstay of our social understanding, but rather as the last resort to which we turn when all the standard mechanisms of social understanding and interpersonal accommodation break down.

In any event, the discussion in this chapter has opened up the possibility of a range of different ways of thinking about the scope of commonsense psychology. At one extreme is the broad construal. Theorists who take this position think that a grasp of commonsense psychology underlies all our social understanding and social interaction – either explicitly, when we are reflectively employing the concepts and categories of commonsense psychology, or implicitly. At the other extreme is the narrow construal, according to which we only employ commonsense psychology on those occasions when we do so explicitly and reflectively. No doubt the truth lies somewhere in the middle.

In exploring the dialectic between these two approaches we have seen that commonsense psychology, so frequently taken to be a unitary phenomenon, does in fact have a complicated articulated structure. Figure 7.2 tries to capture this structure. Broad theorists stress the components on the leftmost side of the diagram. They hold that our social skills and abilities should all be understood in terms of the conceptual framework of propositional attitude psychology – and in particular in terms of the principles an implicitly known theory of mental states and behavior. They accord little, if any, importance to simulation as a tool of social understanding. At the other

extreme narrow theorists see most social interactions and social understanding as underwritten by more primitive mechanisms, of the sort discussed in section 7.5. As Figure 7.2 makes clear, there is room for many intermediate positions. Commonsense psychology is a rich and complex tool that has many strands.

8 From perception to action

The standard view and its critics

- From perception to action
- Cognitive architecture and the standard view
- The distinction between perception and cognition
- Domain-specific reasoning and the massive modularity hypothesis

This chapter turns from the question of how behavior is explained to the question of how behavior is generated. One might well expect the two questions to have related answers. On the standard model of commonsense psychology, there is a reciprocal relation between the explanation of behavior and the generation of behavior. We explain intentional behavior in terms of beliefs and desires because intentional behavior is caused by beliefs and desires. To the extent, therefore, that we have been thinking about alternatives to propositional attitude psychology as a way of explaining behavior, we seem committed at least to taking seriously the possibility that a significant proportion of intentional behavior may not in fact be generated by propositional attitudes in the manner standardly assumed.

Doubts about the role of propositional attitudes in explaining behavior do not in any sense *entail* doubts about the role of propositional attitudes in generating behavior. We spent some time discussing issues of computational tractability in the previous chapter, and it may well be that there are computational reasons why we should not bring the machinery of propositional attitude psychology to bear even on behavior that is causally produced by propositional attitudes. The alternatives to commonsense psychology considered in the previous chapter might simply be pragmatic short cuts for arriving at explanations and predictions of behavior that are good enough for our practical needs. But, on the other hand, they might not be. There is, at the very least, an open question about the springs of action. This is, moreover, a question that has deep implications for how we think about the architecture of cognition, the interface problem and many other issues at the heart of the philosophy of psychology.

The standard view of the route from perception to action involves a linear flow of information from the sensory periphery into the central belief and propositional attitude system, and an equally linear output flow from that system leading directly to action. In between perception and action fall the central belief-fixing and decision-making processes. In essence, perceptions give rise to beliefs that, in combination with desires and other "pro-attitudes", yield actions. This standard view, and the way it is entrenched in some of the pictures of the mind we have been considering, is explored in

section 8.1. Section 8.2 considers some of the implications that the standard view has for how we think about *cognitive architecture* (the mechanisms that are responsible for the implementation of cognition in the brain). The remaining sections of the chapter discuss challenges to the standard view. Section 8.3 considers how sharp the distinction is between perception and cognition, while section 8.4 explores what has come to be known as the *massive modularity hypothesis*.

8.1 From perception to action: the standard view

At a very general level of description all cognitive systems are embedded in their environment. We can see them as picking up information about the immediate environment in the form of perception, and as acting upon the environment in virtue of their needs and/or desires and the information that they possess about the environment. The question that we will be considering in this chapter is how to think about what takes place between perception and action.

In the simplest case there are direct links between receiving information and acting upon the environment. This is the case in reflex behavior, for example. If I pick up information that an object is moving rapidly towards my face (in virtue of its *looming* in my visual field) then I will flinch. Reflexes are responses that are automatic. They can either be *hard-wired* by evolution, or acquired through some process of conditioning. There is an obvious rationale for the existence of such autonomatic responses. It is advantageous to the organism to be able to act immediately when it detects something that might be harmful (or advantageous).

Nor are reflexes the only type of automatic behavior. Ethologists studying animal behavior frequently discuss what they call *innate releasing mechanisms*. These are fixed patterns of behavior that are more complex than reflexes, because they involve chained sequences of movements rather than a simple reaction, and yet that seem to be instinctive (Tinbergen 1951). A good example is the pecking response in herring gull chicks. Newly hatched herring gulls are particularly sensitive to sensory input correlated with the length, movement and coloration of the adult herring gull's bill and when they encounter such input they respond by pecking vigorously at whatever presents the appropriate input (usually, of course, the adult's bill tip). The adult herring gull responds by feeding the chick. Innate releasing mechanisms, such as the herring gull pecking response, have the following characteristic (Lea 1984).

- They are triggered by specific stimuli.
- They always take the same form.
- They occur in all members of the relevant species.
- Their occurrence is largely independent of the individual creature's history.

- Once launched they cannot be varied.
- They have only one function.

In innate releasing mechanisms, just as in simple reflex mechanisms, there is a strict relation between stimulus and response. Once the stimulus (the bill tip of the adult herring gull, for example) is detected, the response (pecking) follows automatically and with a more or less invariant pattern.

The same strict relation between stimulus and response occurs in conditioned behavior. The simplest example of conditioning is *classical conditioning* (also known as *Pavlovian conditioning*). Classical conditioning involves training an organism to respond in a certain way to a given stimulus by associating that stimulus with a further stimulus that the organism finds either attractive or repellent. So, to take a well-known example, a dog might be conditioned to salivate at the sound of a bell by presenting it with food at the same time as a bell is rung. Eventually the dog's natural reaction of salivating in response to the food is provoked simply by the sound of the bell that has become associated with the food. In *instrumental conditioning* the response in question is more complex – typically, an action (such as pressing a lever or going to a particular location in a maze). The action is either reinforced (when it is followed by a reward) or inhibited (when it is followed by a punishment). Instrumental conditioning is a way in which an organism can learn to respond in certain ways to stimuli for which it does not have hard-wired responses.

The processes of classical and instrumental conditioning have been much studied by animal behaviorists because it is thought that they provide the key to understanding those aspects of animal behavior that cannot be understood purely in terms of hard-wired responses and innate releasing mechanisms. There is room for considerable debate about whether some combination of hard-wired responses, classical conditioning and instrumental conditioning can account for all animal behavior. The issue here is really about the scope and range of a particular type of explanation. This is a type of explanation that explains behavior on the assumption of direct links between stimulus and response. These links do not have to be hard-wired (as they are in reflex responses and innate releasing mechanisms). They can be learnt (as in classical and instrumental conditioning). But whether the links are learnt or hard-wired, there will be predictable and automatic connections between particular classes of stimuli and particular classes of response.

Many theorists have assumed (and it is a natural assumption to make) that the belief–desire explanation characteristic of commonsense psychology only comes into the picture when it is *not* possible to employ stimulus-response explanations (Fodor 1986). Psychological explanations of behavior are only necessary when no such input–output links can be identified. They explain behavior in terms of the beliefs that the creature has about its environment, rather than simply in terms of the stimuli that it detects. Beliefs and other propositional attitudes function as intermediaries between

sensory input and behavioral output. This provides a way of explaining behavior based upon the relations between these content-bearing states. This is a style of explanation that exploits the following:

1 The different ways that content-bearing states can be generated on the basis of perceptually derived information.
2 The different ways that content-bearing states can interact with each other.
3 The different ways those combinations of content-bearing states can generate behavior.

This style of explanation is of course what is distinctive about commonsense psychology.

This way of thinking sits very naturally with a particular way of thinking about the route from perception to action, and with a very general and intuitively plausible picture of the organization of cognition (Hurley 1998). On this view we can divide cognition into three stages. The first stage is the input stage, in which information about the environment is picked up and translated into a format suitable for being brought to bear upon the subject's system of propositional attitudes. Let us call this the *perceptual stage*. The second stage might be termed the *central-processing stage*. Here the perceptually derived information is integrated into the subject's propositional attitudes. This process typically begins with the formation of a perceptual belief. The perceptual belief may be a perceptual belief to the effect that a certain goal is attainable, or may impact upon the process of decision-making in some other way. It may make some beliefs more probable and others less probable. It may have implications that are incompatible with existing beliefs. It may equally prompt a process of practical deliberation. It may in fact make available an entirely new goal to the cognitive system. One might think of the central-processing stage of cognition as involving processes of belief fixation and decision-making. This stage of cognition has characteristic types of output, just as it has characteristic types of input. The most typical are intentions to behave in certain ways. This behavior could be communicative behavior, of course, or it may be direct action upon the world. The final stage, in this very schematic account of one plausible way of thinking about the overall organization of cognition, is implementing the output of the central-processing stage by generating the appropriate form of behavior – which might, once again, be speech behavior or motor behavior. Let us call this the *motor stage*.

There are many different ways of working out this general picture of the route from perception to action. The picture of the autonomous mind offers a distinctive way of thinking about what happens in the intermediate stage between perception and action. Autonomy theorists tend to emphasize the role of conscious deliberation in decision-making and belief formation. This is part and parcel of their emphasis on the norm-governed nature of

theoretical and practical reasoning. According to autonomy theorists such as McDowell and Davidson, we are governed by norms in the strong sense that we use those norms to regulate our thoughts and actions. We do not simply conform to norms (for the most part), but actively monitor the extent to which we conform to the normative principles of coherence and rationality. The system of propositional attitudes is viewed by autonomy theorists as a complex inferential structure bound together by logical and probabilistic relations. These logical and probabilistic relations mean that a change somewhere in the system (a new perceptual belief, for example) will have complex ramifications throughout the structure. This picture assumes a high degree of reflective control over belief formation and decision-making. It assumes in addition that a considerable proportion of our practical and theoretical reasoning is consciously accessible. As we saw in Chapter 6, these assumptions lead autonomy theorists to see a radical incommensurability between commonsense psychological explanation (which tracks the rational relations between propositional attitudes and behavior) and the forms of explanation operative lower down in the hierarchy of explanation.

The representational and functional approaches to the mind take a far more down-to-earth view of what goes on between perception and action. The emphasis for philosophical functionalists and representational theorists is on the causal dimension of how perception impacts upon the propositional attitude system; how the propositional attitude system evolves; and how particular combinations of propositional attitudes in particular circumstances give rise to particular actions. Particular patterns of sensory stimulation typically give rise to particular conscious perceptions that in turn have typical effects within the system of propositional attitudes as a whole. Different configurations of propositional attitudes feed into behavior in different ways as a function of the relevant social and physical environment. This stress on causation goes hand in hand with a downplaying of the importance of reflective adherence to normative principles of rationality and consistency. To take a simple example, whereas an autonomy theorist might think that the principle of *modus ponens* operates as a normative principle governing deliberation (to the effect that someone who believes that p and believes that $p \Rightarrow q$ is rationally committed to believing that q – or else to revising one or both of the two original beliefs), a philosophical functionalist would be inclined to hold that it is a causal law that thinkers who believe that p and believe that $p \Rightarrow q$ typically end up either believing that q or revising one of their original beliefs. One way of thinking about the difference here is that, for the autonomy theorist but not for the philosophical functionalist, the thinker's understanding of the principle of *modus ponens* is likely to play a role in explaining how they end up with the belief that q. By the same token, whereas it is typical of philosophical functionalism and the representational theory to take the formation of perceptual beliefs as a brute fact underwritten by perceptual mechanisms, many autonomy theorists think that there are rational connections between perceptual beliefs and the

perceptions on which they are based – and hence that perceptual beliefs can be rationally accountable to perceptions.

Psychological functionalists take a different view of psychological explanation. The aim of psychological functionalism is not to find causal laws explaining how, for example, certain patterns of sensory stimulation give rise to the belief that there is a chair in front of one, but rather to explain the general mechanisms that lead from sensory stimulation to belief formation. This explanation is carried out by a process of *functional decomposition*, breaking the general task down into more specific tasks and showing how those more specific tasks can in turn be broken down into still more specific tasks. Mechanisms are identified to perform individual tasks. As we saw in Chapter 4, the typical result of functional decomposition is what is frequently called a *boxological* account of cognition of the sort that can be illustrated in a flowchart that tracks the flow of information through a succession of mechanisms understood in terms of the task they are performing.

Figure 8.1 is a typical example of such a boxological account – Bruce and Young's functional model of how face processing works. Face processing is widely believed to be a specialized cognitive function, carried out by particular neural circuits dedicated to that task. The functional model of face processing proceeds by breaking the global task of recognizing and identifying faces down into a series of sub-tasks. Some of these sub-tasks are sequential and others are performed in parallel. The first operations performed by

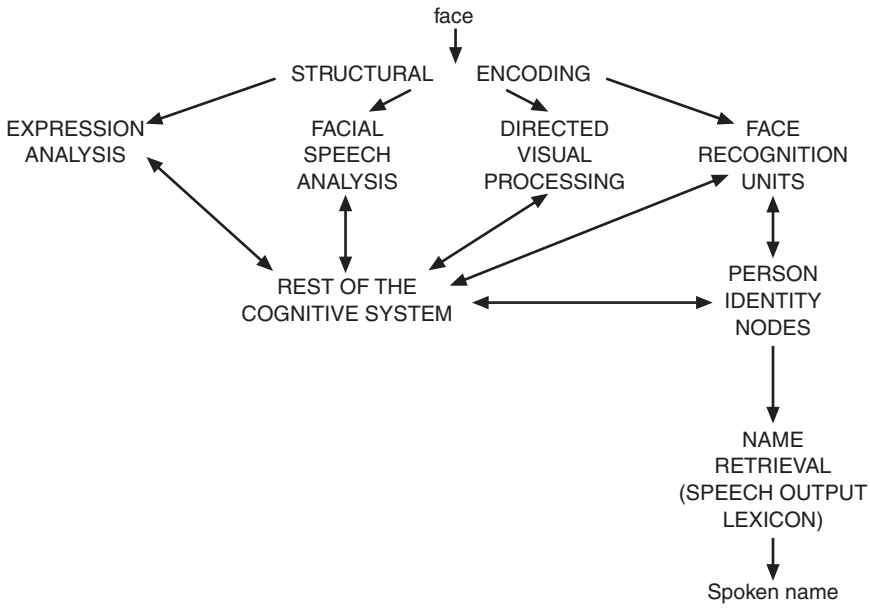

Figure 8.1 Functional model of face processing (source: based on Bruce and Young (1986)).

the face processing system involve straightforward perceptual processing – analyzing the perceived expression and any accompanying vocalizations, as well as scanning the face for any distinctive features. The outputs of these processes feed into an initial face recognition mechanism that itself receives inputs from the central system of propositional attitudes. Output from the face recognition mechanism serves as input to a mechanism that associates the familiar face with a particular person and then to a further system that retrieves the name of that person. The model shows how the initial perceptual information undergoes a series of increasingly complex forms of processing. The identity of the discrete processing sub-tasks is determined by a range of factors, including *dissociations* observed in patients with specific disorders of face processing. So, for example, the existence of *prosopagnosic* patients with relatively unimpaired abilities to perceive faces but who are incapable of putting names to those faces is frequently taken to show that face perception and face recognition are distinct tasks.

The boxological approach shares with the autonomy theory and philosophical functionalism a commitment to the standard view of the route from perception to action. We see, for example, how a clear distinction is made between the perceptual processing involved in face recognition and central processing (labeled in the model as "the rest of the cognitive system") – although there is some two-way traffic between perceptual processing and central processing (most obviously because one would expect background information and expectations to feed into face recognition). The transition from perceptual processing to central processing is made when the system starts to match a familiar face to stored representations of persons (person identity nodes) and names. The further transition from central processing to motor processing comes at the very end of the process, with the utterance of the relevant name.

These very different pictures of the mind share the broadly tripartite conception of three stages of cognitive processing. Although each has rather different views about how cognitive processing works, the general idea that we can distinguish perceptual processing (input processing), central processing and motor processing (output processing) is deeply engrained in philosophical functionalism, psychological functionalism and the representational theory of mind. Things are slightly more complicated when it comes to the picture of the autonomous mind, because autonomy theorists are frequently antipathetic to all talk of information processing, but here too we can identify a commitment to a distinctive type of central processing involving propositional attitudes. The propositional attitude system takes perceptions as inputs and then generates various types of intentions to behave in certain ways (including, of course, verbal behavior). Although this three-stage model of cognition seems highly intuitive, it nonetheless goes hand in hand with a range of more specific commitments for how we think about cognition and the architecture of cognition. We shall explore these in the next section.

8.2 Cognitive architecture and the standard view

At a suitable level of abstraction, proponents of the autonomous, functional and computational minds all think in somewhat similar terms about the route from perception to action. Each picture makes comparable distinctions between what goes on during the perceptual stage, during the central stage and during the motor stage. What happens, though, when we shift our attention from thinking about the very general functions performed at each stage to thinking about the mechanisms that might carry out those functions?

In very broad terms, the three-stage view is a view about how information is processed in the brain. Information ultimately derived from the visual environment undergoes a range of different transformations and operations. These transformations and mechanisms are carried out by different mechanisms. Can we say anything more detailed about the types of transformation and operation that might be involved? Do these operations and transformations depend upon the relevant information being encoded in a particular way? Does the possibility of certain types of information processing, for example, depend upon the information it involves being represented in a language-like form? Can we identify any general properties that are possessed by mechanisms that carry out specific types of information processing?

These are questions about the architecture of cognition. We start to think about the architecture of cognition when we reflect, not just on the particular tasks that might be required in a particular situation (the task, for example, of parsing the visual array so that it is segmented into objects), but on how a cognitive system might carry out that task. Cognitive processing has to start somewhere. There has to be some initial information with which the system has to work – an initial representation, perhaps, of sudden changes in light intensity. Correlatively, cognitive tasks tend to require a determinate type of output – a representation of the boundaries and edges of objects around the perceiver, for example. The basic problem in understanding a cognitive system is to understand how it gets from the input to the output. How can a representation of the sudden changes in light intensity be transformed into a representation of bounded objects? What sort of transformations will be required to move from input to output? How must the initial information be encoded to allow those transformations to take place? The issues here are reminiscent of those that arose when we looked at Marr's three levels of explanation in Chapter 2. Issues of cognitive architecture emerge at what Marr calls the *algorithmic* level – the level at which we move beyond thinking about the general tasks that a system is trying to perform and start thinking about the details of how that task might be effected.

Thinking about cognition in terms of cognitive architecture involves a shift in emphasis from the analysis of cognitive tasks and cognitive functions to the analysis of cognitive mechanisms. We have already looked at the two

dominant models of cognitive architecture in earlier chapters. The picture of the computational mind is based upon a range of hypotheses about the architecture of cognition. The hypothesis upon which we have concentrated so far is the language of thought hypothesis. This is a hypothesis about how information has to be encoded in cognitive systems for certain types of processing to be possible. As we saw in Chapter 4, the language of thought hypothesis is generally motivated as a hypothesis about the way in which information has to be encoded for *central* processing. The computational picture of the mind is linked with a further hypothesis about cognitive architecture – a hypothesis that sharply distinguishes the mechanisms involved in central processing from those involved in perceptual processing and motor processing. This is the *modularity hypothesis*, which will be explored further in this section.[1]

We observed in Chapter 5 that the picture of the neurobiological mind is closely connected with a competing view of cognitive architecture – the approach to cognitive architecture associated with connectionist modeling and artificial neural networks. This way of thinking about the mechanics of mind takes issue with certain fundamental tenets of the computational approach. The connectionist approach places little emphasis on language-like internal representations and does not see central processing as being as clearly demarcated as it is taken to be by the other pictures of the mind. The neurocomputational approach to cognitive architecture can be developed as a counterweight to the standard, three-stage view of the route from perception to action, which sits most easily with the computational picture of the mind. In this section we will be looking primarily at the implications for cognitive architecture of the standard view of the route from perception to action.

The three-stage model requires clear distinctions between, on the one hand, central cognition and perception and, on the other, central cognition and motor control. There are two very natural ways of marking these distinctions. First, the distinctions can be marked at the level of the tasks performed. We might distinguish, for example, the tasks involved in perceptual and motor processing from the tasks involved in central processing. Second, we might mark the relevant distinctions at the level of the mechanisms that carry out those tasks. Of course, thinking about tasks and thinking about mechanisms are closely related; because an important part of what distinguishes different mechanisms are the respective tasks that they perform.

Psychologists and cognitive scientists often draw a sharp distinction between different types of cognitive task – and in particular between low-level cognitive tasks that are essentially perceptual and high-level tasks that are essentially central. It is easy to think of paradigms of the two types of task. We can see them exemplified in the boxological diagram of face

1 We will return to the language of thought hypothesis in the next chapter.

perception. The tasks carried out in the initial processes of expression analysis and speech analysis are clearly at the perceptual end of the spectrum, while those implicated in the processes of identifying a given face are clearly at the central end of the spectrum. Although it is not quite so clear how to make a principled distinction between the two types of task, there are several potential candidates. One criterion that might distinguish complex cognitive tasks from simple cognitive tasks is the involvement of memories. Analyzing facial expression involves analyzing the features of a perceived face in line with some form of relatively primitive categorization (determining, for example, whether the face is sad or angry). As we will see further below, this can be described as a process of template matching. But it does not require matching perceived faces to previously seen faces. Nor does it involve drawing upon background knowledge and expectations. Another related criterion is the involvement of some type of reasoning. At some level identifying faces can require forming hypotheses and thinking about how plausible they are (how likely is it that *X* will be here in the supermarket? Could that really be *Y*? I thought she was on holiday). No such reasoning is required to determine whether someone has black eyes or blue eyes, or whether they look happy or sad.

Similar points can be made about the distinction between central processing and motor processing. Practical decision-making is a paradigmatic central process. It clearly involves reasoning – weighing up the advantages and disadvantages of the different possible courses of action; working out what the different potential outcomes of each course of action might be; thinking about how likely and how desirable those outcomes might be; working out whether any of the different possible courses of action are incompatible with any deeply held principles or prohibitions; and so on. This practical reasoning will involve taking into account a wide range of background information and stored knowledge. However, once the particular course of action has been decided upon, the remaining processing is merely implementational, simply a matter of calibrating body movements to achieve the desired result. There is no further reasoning involved and no need to involve stored knowledge or background information. The processes of executing a particular motor behavior are radically different from the processes of planning that behavior.

The two different types of cognitive task that we have identified will require different types of processing to implement them – and it is natural to think that these different types of processing will be carried out by different types of cognitive mechanism. One might appeal at this point to the distinction between modular, peripheral processes and non-modular, central processes. The thought here is that modular processes are responsible for processing perceptual input and for controlling motor behavior, while belief formation and decision-making are carried out by non-modular processes. The distinction between modular and non-modular brings with it a range of further distinctions. If processing in the perceptual and motor stages were

modular, then one would expect it to be carried out by mechanisms that are *domain-specific* (dedicated to performing a highly specialized type of task) and *informationally encapsulated* (unaffected by what is going on in other specialized modules and in central processing). Central processing, in contrast, to the extent that it is non-modular is *domain-general* and able to draw upon all types of information that are centrally stored as well as upon the outputs of all the peripheral modules that feed into central processing.

We see, therefore, the outline of a very influential picture of the route from perception to action. The process begins at the sensory periphery, with the operation of *transducers* that convert sensory stimulation (patterns of irradiation, or sound-waves) into a format that can be used by subsequent stages of processing. These transducers feed into the modular systems responsible for the earliest stages of perceptual processing. In the case of vision, for example, these early systems might detect sudden changes in light intensity (what Marr called *zero-crossings*). The outputs of these modular systems then serve as inputs into the next stage of modular processing – a stage, in which, for example, information about zero-crossings and other such low-level features is used to construct a sketch of the edges and boundaries in the perceived environment. Successive stages of modular processing eventually lead to a representation of the distal environment as containing 3D objects standing in determinate relations to the perceiver and to each other. This representation of the environment marks the limit of modular processing.

The various stages of modular visual processing generate a representation of the environment that can be used for object identification and object recognition. The processes of object recognition and object identification are not themselves modular. Modular processing is *data-driven* and *bottom-up*. It involves carrying out a limited range of operations on data that are provided by the immediately preceding stage of modular processing. Object recognition and object identification, on the other hand, require exploiting background information, memories and potentially very broad categories of general knowledge. We might describe these processes in the jargon of computational psychology as establishing and exploiting an interface between modular and non-modular processes. Or we might describe them in more standard philosophical jargon as generating conscious perceptions of the environment. These conscious perceptions of the environment in turn lead to the formation of perceptual beliefs. In some cases these perceptual beliefs simply involve taking perceptions "at face value". In other cases, however, things are more complicated. Perceptions are often misleading and need to be corrected. Information from one modality needs to be calibrated with information from other modalities. Potential conflicts with other beliefs need to be taken into account and accommodated, in one way or another. Even the process of forming perceptual beliefs, therefore, requires a degree of *top-down* central processing.

Once perceptual beliefs have been formed, the way is clear for general

inference about the implications for action and reaction of the distal environment as represented. Practical decision-making can come into play. Once again the full range of general knowledge, preferences, memories and plans for the future is potentially relevant. Jerry Fodor, whose 1983 book *The Modularity of Mind* put the modular/non-modular distinction on the map, has coined two useful words to capture what is distinctive about central processing. Central processing, he suggests, is *Quinean* and *isotropic*. What he means by describing the propositional attitude system as Quinean is that it has certain vital epistemic properties defined over the system as a whole. It is the system as a whole that is evaluated for consistency and coherence, for example. We cannot consider the truth or falsity of individual beliefs in isolation, because our attitude to them will be a function of how we think about other elements of the system in which they are embedded. The isotropic nature of central processing is in many ways a corollary of its Quinean property. The propositional attitude system forms a holistic structure in such a way that any member of that system is potentially relevant to the confirmation of any other.

Clearly, therefore, the decision-making, planning and belief-fixation carried out by central processing cannot be sequentially understood in the manner of modular processing. Sequential processing returns, however, with the modular output systems that translate the decisions and intentions determined centrally into motor behavior. The route from intention to execution involves many steps. This can be seen even when we think about the simplest kind of motor behavior, namely, a simple reaching movement. Suppose that the end result of the central decision-making processes is an intention to reach out to pick up a glass of water. The successful execution of this movement requires a complex calibration of self-specifying information with information about the environment. The position of the glass relative to the arm that will be doing the reaching needs to be calculated. This initial calculation involves calibrating information about the arm (where it is relative to the rest of the body) with information about the glass. This calibration will involve coding the location of the glass on a coordinate system that is centered on the hand (as opposed to the coordinate system centered on the eyes in which visual information about the location of the glass is given), in addition to bringing this spatial information in line with information derived from muscle sensors and other forms of *proprioception* about the location of the arm. The aim here is to construct a representation of the space around the agent that incorporates both the starting-point and the end-point of the projected movement. Once this has been achieved the next step is to calculate a trajectory that will lead from the starting-point to the end-point. This trajectory will initially be calculated in kinematic terms (that is to say, in terms of the sequence of positions that the arm will occupy). The next stage is to calculate an appropriate combination of muscle forces and joint angles that will take the arm along the appropriate trajectory. But the movement of the arm is still only one element in the reaching

movement. The fingers need to be positioned at the right aperture to grip the glass, and beginning of the grasping movement needs to be timed to coincide with the arm's arrival at the glass. Even once the glass has been successfully grasped, it needs to be brought back to one's lips. This involves a comparable set of calculations that cannot simply be carried out by reversing the earlier movements. Not only is the end-point of the "return" movement completely different from the starting-point of the "outward" movement, but the forces involved will be completely different due to the additional weight of the glass.

When the route from intention to execution is described in these terms it is natural to think that it will be carried out by a series of modular processes. A modular architecture seems well suited to carrying out the highly specific tasks involved at each stage in the process. It seems plausible, for example, that the systems translating retinotopic (eye-centered) coordinate frames into arm-centered coordinate frames are *informationally encapsulated*. They do not need to draw upon background knowledge or memories. Nor do they need access to the outputs of any other processes, except those that yield as output the relevant retinotopic coordinates. By the same token these and the other mechanisms involved in executing action are *fast*. Each stage of the planning process needs to be completed before the next stage can begin – and there are many such stages that need to be completed before, for example, the prey runs away or someone else picks up the last glass on the tray. Nor would one expect these mechanisms to be domain-general. The mechanism that computes a kinematic trajectory through the space immediately surrounding the body is unlikely to serve also to compute the kinematic trajectory that I need to take to drive from Detroit to Philadelphia.

It seems, therefore, that the standard view of the route from perception to action sits easily with a conception of the architecture of cognition that sees non-modular central processing as sandwiched between, on the one hand, a range of modular processes that provide perceptual input into the central processes of belief fixation and decision-making and, on the other, a range of modular processes that control the motor output from those central processes. The cognitive mechanisms that carry out the information processing involved in the perceptual and motor stages are fundamentally different from those involved in central processing. Whereas central processing is domain-general, perceptual and motor processing is domain-specific. Central processing is Quinean and isotropic, whereas perceptual and motor processing is highly specialized.

The distinction between modular and non-modular cognitive systems goes naturally with the view that central processing is the realm of the propositional attitudes. This is clearly the case for belief fixation and decision-making. But even relatively simple central operations such as object recognition are influenced by propositional attitudes. In fact, one obvious consequence of the claim that central processing is isotropic is that central processing turns out to be *permeated* by the propositional attitudes.

This has a further, and less frequently stressed, consequence. It follows that any cognitive task or cognitive mechanism that cannot be understood in modular terms in some sense involves propositional attitudes – either directly or indirectly. As soon, therefore, as we leave the domain of modular processes we enter the realm of beliefs and desires.

This way of thinking about cognitive architecture complements what in the previous chapter I termed the broad construal of the scope of commonsense psychology. According to the broad construal of commonsense psychology, the machinery of propositional attitude psychology is our principal tool for making sense of the behavior of other people. The basic idea is that whenever we are confronted with behavior that cannot be understood as an immediate response to obviously identifiable features of the environment we try to make sense of that behavior by identifying the particular configuration of beliefs and desires from which it might have emerged. One way of justifying this practice would be to say that all behavior that cannot be understood as an immediate response to obviously identifiable features of the environment *ipso facto* involves central processing – and therefore is permeated by propositional attitudes in such a way that one has no hope of understanding it without bringing in the machinery of propositional attitude psychology. Generally speaking, it seems plausible that, for any theorist committed to the distinction between modular and non-modular processes, the greater stress they place on the role of the propositional attitudes in non-modular central processing, the more inclined they will be to a broad construal of the scope of propositional attitude psychology.

Although the distinction between modular and non-modular processing, together with the three-stage view of the route from perception to action, plays a dominant role in the scientific study of the mind, as well as in the pictures of the functional, autonomous and computational mind, it is not the only way of looking at the structure and organization of cognition. In the remainder of this chapter we look at some ways of placing pressure on the standard view. Some of these alternatives to the standard view are very closely related to the alternative way of thinking about cognitive architecture put forward by proponents of the neurobiological mind, but others have more general scientific and philosophical origins.

8.3 The distinction between perception and cognition

The standard view of the route from perception to action is closely tied to the possibility of making a clear distinction between perceptual processing and central processing. As we saw in the previous section, there are various ways of marking this distinction. Central processing can be characterized by the involvement of memories, by the need for some form of reasoning, by the integration of background knowledge and expectations, or by being relatively slow and non-specialized. Perceptual processing, on the other hand, does not involve memories, reasoning, background knowledge or

expectations and is carried out by processes that are relatively fast and specialized. This distinction at the level of cognitive architecture goes hand in hand with a higher-level distinction between different types of cognitive task – between perceptual tasks and central tasks. So, one very natural question to ask is whether the two distinctions do indeed map onto each other. Is it in fact the case that all the tasks that we would think of as tasks of central cognition are best viewed as performed by central processes? Are there cases where what seem to be high-level central tasks are in fact carried out by relatively low-level perceptual processes?

Let us start with the thought that much perceptual processing involves pattern recognition. Of course, not all types of pattern recognition are straightforward. Think, for example, of the complex forms of pattern recognition involved in understanding a mathematical proof, or finding one's way around a new city. Recognizing patterns and structural similarities between different phenomena can involve drawing upon a wide range of background knowledge and require lengthy processes of conscious deliberation before the final "flash" of insight. This is perceptual processing only in a purely metaphorical sense (the sense in which one might say that one suddenly "sees" the solution to a problem). But, within the general category of pattern recognition, we can identify a narrower sub-category that does seem to be purely perceptual. We can term this template matching, where the task is simply to work out whether or not a particular pattern matches a particular prototype or template. Identifying a shape as a triangle or a letter as a 'p' are good examples. Template matching counts as perceptual according to the criteria we have discussed. It does not involve reasoning – in fact, template matching is often taken to be the antithesis of reasoning. Nor does it require integrating background knowledge or expectations. And it is both fast and specialized (since any cognitive mechanism involved in template recognition will have only a limited number of available templates or prototypes).

Suppose we take template matching to be paradigmatic of perceptual processing. It certainly seems to be the case that many of what we would intuitively take to be perceptual tasks can in fact be carried out by processes of template matching. Template-matching is deeply implicated in perceptual processing – from the initial stages in which what matters is matching changes in light intensity to the template for an object boundary, to the final stages in which segmented elements of the visual field are matched to templates for different shapes. With this characterization of perceptual processing in mind, one obvious potential difficulty for the standard view of the route from perception to action would be examples of central tasks that can be understood in terms of template-matching – that is to say, instances of the type of cognitive task that we would intuitively think as involving propositional attitudes, reasoning, and so forth but that can be seen as involving the matching of perceived situations to templates or prototypes.

We saw two different candidates for such cognitive tasks in the previous chapter. The first candidate comes from the application of the simple heuris-

tics and rules of thumb that I proposed play a significant role in social cognition and social interaction. The example we looked at was the TIT-FOR-TAT rule in social situations that can be modeled as repeated prisoner's dilemmas. The TIT-FOR-TAT rule states that one should start out in any such social exchange by cooperating, and then do whatever the other participant does (cooperate if she cooperates, and defect if she defects). In applying such a rule the main thing the agent has to do is identify what the other participant did in the previous round – that is, to identify whether they are dealing with a defector or a cooperator. It seems plausible to view this as a process of template matching – of fitting the appropriate behavior to one's prototype either of cooperation or of defection.[2]

The second candidate for a central cognitive task carried out by template matching emerges from the hypotheses that a surprisingly large amount of our social interaction is carried out by means of social scripts and routines. Whereas it is usual for philosophers to think that social coordination requires forming beliefs about other people's beliefs, desires and intentions we considered the possibility that many social interactions are sufficiently standardized to be successfully negotiated by identifying the relevant social roles and acting accordingly. Once the relevant social roles are identified the script takes over and the interaction runs according to rule. But how are the social roles identified? How do we know when to initiate the restaurant script, or the dry-cleaners script? This might well be viewed as an instance of template-matching – of matching the perceived behavior to one or other of the templates associated with the social scripts and routines into which one has been enculturated.

Here, then, are two cases of apparently central processes that might be seen as carried out by processes of template-matching that fall on the side of perceptual processing rather than central processing. If the suggestions in the previous chapter about the significance in social cognition of social heuristics and social routines are correct, then an important element of social cognition is hard to fit into the standard view of the route of perception to action. At the level of task analysis what is going on clearly seems to be part of central cognition. At the processing level, however, the mechanisms responsible seem more akin to perceptual mechanisms.

This general idea that much of what we intuitively think of as central processing can be carried out by mechanisms of template matching has been strenuously advocated by proponents of the neurocomputational mind (e.g. Churchland 1989b). It is no accident that artificial neural networks, which as we saw in Chapter 5 offer perhaps the most promising way of modeling the neural basis of cognition, are particularly effective on tasks that involve pattern recognition and template matching – as emerged very clearly in the

2 The example of TIT-FOR-TAT is very important in the massive modularity hypothesis. We will come back to it in section 8.4.

example of the rock/mine detector we considered earlier.[3] In fact, since most neural network models have employed forms of supervised learning algorithms (on which the network weights are modified as a function of how closely the network output approximates to the desired output), it seems plausible to describe neural network models as mechanisms of pattern recognition. The mainstay of the neurocomputational approach to the mind is the idea of a co-evolutionary research ideology – the idea that our thinking about personal-level cognition and subpersonal cognitive architecture can be informed by theorizing about the neural implementation of cognition, and vice versa. The extensive evidence from neural network modeling that neural networks can carry out sophisticated cognitive tasks might therefore be taken (and indeed has been taken) as evidence that the line between perception and cognition is far less sharp than it is standardly taken to be.

Pressure might be put on the sharpness of the distinction between perception and cognition from another direction. Suppose we think of central cognition as the domain of the propositional attitudes, where belief fixation and decision-making take place. Central cognition is the name for the way in which we make sense of the world (both the natural world and the social world) and organize our responses to it. Much of this mental activity is conscious (although of course much of it also goes on unconsciously). We frequently reflect consciously about how to make sense of puzzling features of the world, particularly the social world – and still more frequently about how to respond to it. Suppose we ask what the raw material is for this process of conscious reflection. In one obvious sense this raw material is provided by the content of perception – by how we perceive the physical environment and the people around us. As was suggested in section 8.1, we form perceptual beliefs by integrating how the world appears to us perceptually with our beliefs and other information about the world. Sometimes these perceptual beliefs simply endorse the content of perception, when we believe that the world really is the way it appears to be. On other occasions things are more complex and we make adjustments for ways in which we know that perception can be misleading, as in calculating distances or identifying colors in poor light. But perception is essentially the point of contact between the belief system and the environment. Conscious perceptions are the input to the propositional attitude system, when we think about cognition from the personal level. This raises the question of how the conscious perceptions that serve as input into the propositional attitude system relate to the outputs of modular perceptual processing. Clearly, if we are to have a good fit between our personal-level account of cognition and our account of cognitive architecture, then these two ways of thinking about conscious perceptions must map onto each other. But things are far from straightforward in this area.

3 Many other examples can be found in McLeod *et al.* (1998). See also Bechtel and Abrahamsen (1991).

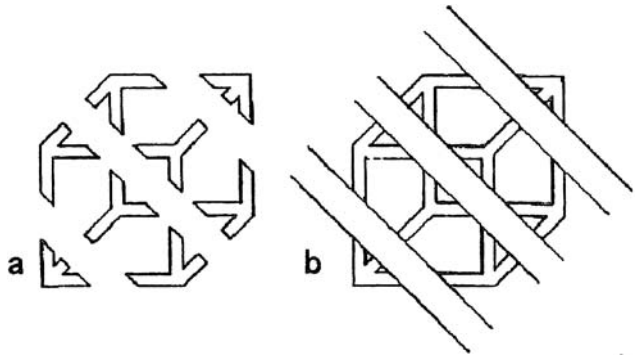

Figure 8.2 Occlusion and mid-level vision (source: adapted from Kanizsa (1979) in Wilson and Keil (1999, p. 545)).

Theorists of vision often make a distinction between three different levels of visual processing (see, for example, the Preface to Ullman 1996, Nakayama 1999, and Peterson 1999). Marr's theory of vision, which we looked at in Chapter 2 and to which we have returned on a number of occasions, is generally described as a theory of *early visual processing*. The task of early visual processing is to derive a three-dimensional representation of the shape and spatial arrangement of a distal object in a form that will allow that object to be recognized. But this is only the first stage in thinking about what the visual system as a whole has to do. In order to appreciate a further set of processing tasks that the visual system needs to carry out we can consider the following two figures, derived from the work of the Gestalt psychologist Gaetano Kanizsa (Kanizsa 1979).

Figure 8.2, on the left, shows us an apparently disconnected set of fragments. The figure on the right shows us exactly the same fragments, but with three superimposed diagonal lines that allow the fragments to be seen as components of the familiar Necker cube. It is standardly thought to be the job of mid-level vision to impose the type of order upon perception that distinguishes the figure on the right from the figure on the left. An important part of this job is to make sense of situations where (as in the Necker cube illustration) one object is partially occluded by another. Another part of the job is to cope with different effects created by reflected light and shadows. Figure 8.3 is a good example. The darker figure on the right looks much more two-dimensional than the figure on the left. Accordingly it is more natural to interpret the figure on the right as a silhouette (in fact, a silhouette of a saxophone player). The diminished contrast between figure and ground in the figure on the left, however, means that the dark regions are most naturally seen as shadows (and the figure as a whole as the image of a woman's face).

Figure 8.3 Figure and ground (source: adapted from Kanizsa (1979) in Wilson and Keil (1999, p. 546)).

As the second pair of figures brings out, the processing involved in mid-level processing is a prerequisite for object classification and identification. With this we come to the domain of high-level vision. High-level vision is standardly taken to involve influences from memory, context, background knowledge and intention. To take a final and well-known visual phenomenon, it is the job of high-level vision to disambiguate the duck/rabbit image. The processing involved in high-level vision seems clearly to involve top-down processing. It must, therefore, involve central processes.

Let us return, then, to our original question. How do the conscious perceptions that serve as input into the propositional attitude system relate to the outputs of modular perceptual processing? Or, to put it another way, where does consciousness enter the picture in visual processing? At what stage in visual processing can we locate the inputs into the propositional attitude system? It seems clear that early visual processing is quite simply too early. The principal information emerging from early visual processing is information about the shape of objects and the volume of space that they occupy. It is very unclear, however, that there is a level of conscious perception at which we perceive the environment in this way. We do not perceive colored expanses and shapes. We have a choice, then, between mid-level vision and high-level vision.

This is where the potential difficulty arises. What we are looking for is a level of visual processing that can plausibly be viewed *both* as modular *and* as

yielding outputs that can themselves be inputs into the propositional attitude system. That is, we are looking for modular systems that can yield conscious perceptions as output. So, there are two constraints in operation. The first constraint directs us towards the modular components of visual processing, while the second directs us to the *phenomenology* of visual experience. The first constraint leads us to ask: which level of visual processing is modular? The second leads us to ask: which level of visual processing best captures how the world appears to us in visual perception? The problem is that the two constraints may well pull us in different directions. The modularity constraint leads us to mid-level vision, as opposed to high-level vision. The top-down components of high-level vision seem to make it too dependent upon central-processing to count as modular in the sense that we need (recall that we are looking for a way of thinking about visual perception that allows a sharp distinction between perception and cognition). Yet it will seem plausible to many that the phenomenology of visual experience is best captured by the high-level account rather than by the mid-level account.[4]

The point is sometimes put by saying that perception is concept-laden – that there is no level of perception where we perceive the world in a manner independent of how we think about it. We do not apply concepts to what we perceive, but rather what we perceive is structured by our concepts and, more generally, by our beliefs about the world. But some philosophers have suggested that there are further ways in which the content of perception can be meaningful that are not necessarily tied to our conceptual capacities (e.g. Peacocke 1992, Chapter 3). So, for example, when we look at objects part of what we see is the object's distance from us. This distance is not given in terms of any standard unit of measure. We do not see objects as a certain number of centimeters away from us. Rather, we see the distance of objects in terms of our own capacities for action. Whether objects are within reach or out of reach their distance from us is presented in terms of the movements that we would have to make to get to them. In fact, we see objects more generally in terms of what we can do with – in terms of the possibilities that they *afford* (Gibson 1979). We see chairs as things upon which we can sit, cups as things that we can grasp, fruit as something that can be eaten. Our perception of the world is imbued with the *affordances* that the environment presents. Again, it may well be that this type of information is neither purely perceptual nor the type of information that can be provided by modular processing.

We see, then, that the standard view rests upon a clear distinction between perceptual processing and central processing that can be questioned in several respects. It is far from clear, for example, that everything that is standardly characterized as a central process actually involves the propositional attitudes. Much of our social coordination may rest upon proto-perceptual processes of matching perceived situations to social templates and

4 This has been contested. Some authors have suggested that the outputs of mid-level vision are accessible to conscious awareness (Jackendorff 1987; Nakayama *et al.* 1995).

prototypes. And there are important questions to be asked about how the subpersonal organization of visual processing maps on to the phenomenology of visual experience. The essence of the standard view is to map the personal-level distinction between perception and cognition onto the subpersonal distinction between modular perceptual processing and non-modular central processing, but it may well turn out that there is no clear-cut mapping in this area.

8.4 Domain-specific reasoning and the massive modularity hypothesis

The previous section considered some ways of placing pressure both on the idea that there is a sharp distinction between central and peripheral processing and on some standard ways of thinking about that distinction. This section explores how reasoning is understood on the standard view of the route from perception to action and considers a far more wide-ranging attack on the conception of cognitive architecture that goes with the standard view.

Let us think back to the basic distinction between perceptual and motor processing, on the one hand, and central processing on the other. We have, first, types of cognitive task that are relatively specialized, that need to derive roughly similar types of output from roughly similar types of input, and that do not seem to depend upon the results of cognitive tasks other than those involved in providing the appropriate sort of input. These are to be distinguished from types of cognitive task, such as belief fixation and decision-making, for which both input and output can vary enormously and to which just about anything can be potentially relevant. At the subpersonal level of cognitive architecture these tasks are supposed to be effected by very different types of mechanism. Domain-specific and encapsulated modules perform the specialized tasks, while the central tasks are performed by mechanisms that are domain-general and able to integrate a wide range of information.

What makes it the case that central mechanisms are able to integrate this wide range of information? The standard answer is that central mechanisms involve forms of inferential transition that are fundamentally different from those involved in peripheral mechanisms. Suppose we define an inferential transition in rather loose terms as a rule-governed transformation from one representation to another. In this sense of 'inferential transition' even peripheral processing involves inferential transitions. Consider, for example, Marr's theory of vision, as a paradigm account of a complex form of specialized processing. Marr's theory postulates a series of representations (or what he calls *sketches*) each carrying increasingly explicit and articulated information about the distal environment. The initial image, for example, carries information only about intensity – the sole variation possible is in the intensity values at different points in the image. The first stage of visual pro-

cessing involves computing the so-called *primal sketch*, which carries information about the geometrical distribution and local organization of changes in intensity values (a necessary first step in identifying contours and shape boundaries) (Figure 8.4).

The transition from image to primal sketch is clearly an inferential transition in the above sense – that is to say, it is governed by rules that allow patterns to be picked out in the overall distribution of intensity values. Yet these rules are highly specialized and domain-specific. They can only operate on a determinate type of input (namely, an overall distribution of intensity values) and they can only yield an equally determinate type of output (namely, a representation of local changes in intensity values).

Now consider the sort of inferential transitions that we regularly perform in daily life. We might, for example, make the following inference: "If that's the cathedral, then the library must be over there. But it's not. So, that can't be the cathedral." Here too we have a rule-governed transition between representations, which in this case are sentences rather than imagistic representations. The rule in question is known as *modus tollens*. This is the rule stating that a conditional (*If A then B*) and the negation of the consequent of that conditional (*not-B*) jointly entail the negation of the antecedent of that conditional (*not-A*). In our example the sentence "that's the cathedral" takes the place of *A* (the antecedent of the conditional) and "the library must be over there" takes the place of *B* (the consequent of the conditional). What is distinctive about this sort of inference is that it makes no difference what sentences one puts in place of *A* and *B*. Whatever one puts in place of *A* and *B* the inference from *If A then B* and *not-B* to *not-A* will always be valid, simply because it is impossible for the two premises *If A then B* and *not-B* to be true and the conclusion *not-A* to be false.[5] Of course, this is another way of saying that this inferential transition is domain-general.

The inference rule of *modus tollens* is a rule of what is known as the propositional calculus (the branch of logic that deals with logical relations between propositions holding in virtue of their truth-values). All the other familiar rules of the propositional calculus are equally domain-general. So too are the rules of the predicate calculus, which is the branch of logic that deals with inferential relations between propositions that exploit the internal structure of those propositions. An example of such a rule might be the rule of existential generalization, according to which a sentence of the form *a* is *F* (where '*a*' is a proper name picking out an individual) entails the existential generalization $\exists x\ Fx$ (namely, there is something that is *F*). Again, it does not matter which individual the name '*a*' picks out or what property '*F*' picks out. The inference is valid because it is impossible for the premise to be true and the conclusion false.

5 This can easily be seen. If the conclusion is false then *A* must be true. Hence, given *If A then B*, *B* must be true. But then the premises cannot both be true.

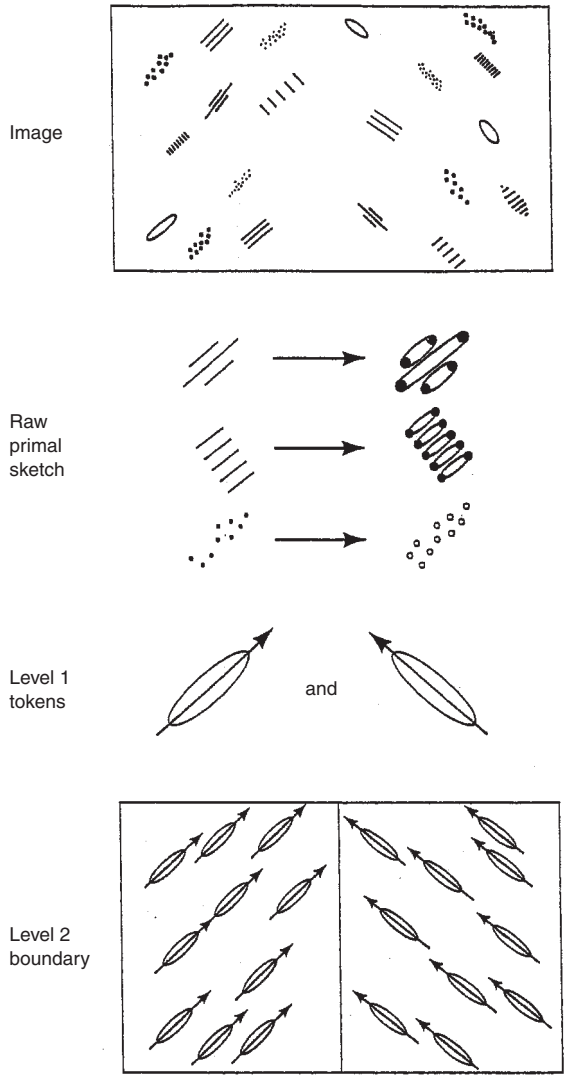

Figure 8.4 A diagrammatic representation of the descriptions of an image at different scales which together constitute the primal sketch. At the lowest level, the raw primal sketch faithfully follows the intensity changes and also represents terminations, denoted here by filled circles. At the next level, oriented tokens are formed for the groups in the image. At the next level, the difference in orientations of the groups in the two halves of the image causes a boundary to be constructed between them. The complexity of the primal sketch depends upon the degree to which the image is organized at the different scales (source: Marr (1982, p. 53)).

The rules of the probability calculus have a similar feature. Suppose we take probability in the subjective sense, so that the numerical probability assigned to a particular proposition is understood to reflect one's personal degree of confidence in that proposition.[6] Then the rules governing the calculations one can perform with that number are completely independent of what the proposition is. It does not matter whether one assigns a probability of 0.25 to the proposition that the next toss of two coins will result in two heads, or to the proposition that the moon is made of green cheese, the probability calculus still dictates that one should assign a probability of 0.75 to the negation of that proposition. Once again we have rules that are domain-general, although the reasons for domain-generality are rather different in this case. Whereas inferential transitions underwritten by the rules of the propositional and predicate calculi are domain-general because they are defined only over formal and syntactic features of sentences, the rules of the probability calculus are domain-general because the proposition to which a numerical probability has been assigned completely drops out of the picture once the numerical probability is in play.

So, we see how a powerful set of ideas comes together. Central processing is the domain of the propositional attitudes. Propositional attitudes, as their name suggests, are attitudes to propositions. Central processing involves plotting the inferential connections between propositions. Some of those inferential transitions are deductive. Others are probabilistic. The types of reasoning that this involves must be domain-general, because it is characteristic of central processing that any belief might be potentially relevant to any other belief, and similarly that all sorts of desires and pro-attitudes might need to be taken into account in working out what to do in a given situation. It looks very much, therefore, as if this overall picture of central processing is closely bound up with the idea that the reasoning involved in belief-fixation and practical decision-making is domain-general in precisely the manner exemplified by the inferential transitions of the propositional, predicate and probability calculi.

This picture of reasoning has come under considerable pressure, however, from both empirical and theoretical directions. From an empirical point of view, a number of studies of the psychology of reasoning have been taken to show that we do not generally employ domain-general inferential transitions. Instead, we are very sensitive to the particular content of the beliefs we consider in ways that suggest that we are employing inferential rules that are highly domain-specific. These empirical studies have been taken up by an influential school of evolutionary psychologists, who have embedded them within a wide-ranging account of cognitive architecture that strikes

6 This is the standard way probability is understood in decision theory, which is the most relevant use of probability theory when we are dealing with processes of belief fixation and practical decision-making. For a brief and non-technical introduction to some of the basic ideas of decision theory, see Allingham (2002).

hard at the standard view of the route from perception to action (Cosmides 1989; Cosmides and Tooby 1992; Pinker 1997). According to proponents of the *massive modularity hypothesis*, there is no such thing as central processing in the way it is standardly understood. There are no domain-general rules of inference that range over the entire domain of propositional attitudes. Instead, we should view the mind as made up of a large number of specialized modules that evolved to deal with highly specific problems confronted by our hominid and pre-hominid ancestors. These specialized modules are domain-specific and employ inferential transitions that are specialized for the relevant domains. Evolutionary psychologists postulate the existence of *Darwinian modules* governing different types of social interaction; our everyday understanding of number; our naïve physics (namely, our understanding of the dynamic and kinematic properties of ordinary objects); our naïve biology (namely, our understanding of the basic properties of living things), and so on. According to the massive modularity hypothesis, the mind is a complex structure of superimposed Darwinian modules. The consequences for the standard view of the route from perception to action are clear. There is no such thing as domain-general central processing at all. Instead of a picture of domain-specific encapsulated modules feeding into a domain-general propositional attitude system the massive modularity hypothesis views the route from perception to action as proceeding via a series of overlapping modules of varying degrees of specialization and domain-specificity.

Let us begin with the experimental studies that have been taken to count against the psychological reality of domain-general reasoning principles.[7] Like much work in the psychology of reasoning, fairly wide-ranging conclusions have been drawn from a relatively small number of basic experimental paradigms (albeit ones that have been refined in a variety of ways). The most influential and best-known experiments in the reasoning literature are on what is known as conditional reasoning, namely, reasoning that employs the "if …, then …" construction that many philosophers and some logicians have taken to be the natural language equivalent of the material conditional in the propositional calculus (often symbolized by a '⊃'). What appears to have emerged from extensive research into conditional reasoning is that people are generally not very adept at mastering conditionals in the ways prescribed by the propositional calculus. Experimental subjects consistently fail to be able to apply some fairly fundamental rules of inference governing the conditional. They have particular difficulties with the rule of *modus tollens* that we considered briefly earlier in this section. This is the rule that that a conditional (*If A then B*) and the negation of the consequent of that conditional (*not-B*) jointly entail the negation of the antecedent of that conditional (*not-A*). Moreover, they regularly commit fallacious inferences involving the conditional – fallacies such as the fallacy of affirming the con-

7 The argument from reasoning studies to the massive modularity hypothesis is surveyed and discussed in more detail in Samuels *et al.* (1999).

sequent. To affirm the consequent is to conclude *A* from a conditional *If A then B* and its consequent *B*. We can compare the two forms of inference side by side:

Valid		Invalid	
	If *A* then *B*		If *A* then *B*
	Not-*B*		*B*
	———		———
	Not-*A*		*A*

The two forms of inference are superficially very similar – but in the case of affirming the consequent, as is not the case with *modus tollens*, it is possible to have true premises and a false conclusion.

One well-known study produced striking results, effectively showing that the only basic conditional inference that subjects seem to be able to recognize and apply is *modus ponens* – namely, the rule that *If A then B* and *A* jointly entail *B*. So, for example, in one study 43 percent of the subjects failed to see that *modus tollens* inferences are always valid (Rips 1983). In the present context, however, what is interesting about the psychology of conditional reasoning is not that people seem regularly to commit various types of fallacy. What is significant is that the standard of correctness seems to vary according to the subject-matter. There are particular ways of framing the conditional reasoning tasks on which success rates improve dramatically. This has struck many theorists as suggesting that subjects may not be applying domain-general reasoning principles at all. The most developed studies of conditional reasoning that show this effect are the many variations that have been developed of the so-called Wason selection task.

Let us start with a typical version of the basic task that inspired the whole research program. Subjects were shown the four cards illustrated below and told that each card has a letter on one side and a number on the other. Half of each card was obscured and the subjects were asked which cards they would have to turn over to determine whether the following conditional is false:

If a card has a vowel on one side then it has an even number on the other.

It is obvious that the *E* card will have to be turned over. Since the card has a vowel on one side, the conditional will obviously be false if it has an odd number on the other side. Most subjects get this correct. It is fairly obvious

that the second card does not need to be turned over, and relatively few subjects think that it does need to be turned over. The problems arise with the two numbered cards. Reflection shows (or should show!) that the *4* card does not need to be turned over, because the conditional would not be disconfirmed by finding a consonant on the other side. The conditional is perfectly compatible with there being cards that have a consonant on one side and an even number on the other. The *5* card, however, does need to be turned over, because the conditional will have to be rejected if it has a vowel on the other side (this would be a situation in which we have a card with a vowel on one side, but no even number on the other). Unfortunately, almost nobody sees that the *5* card needs to be turned over, while the vast majority of subjects think that the *4* card needs to be turned over. In one study that gave an analogous reasoning task to 128 college students, for example, only five correctly identified the *E* card and the *5* card as the ones that needed to be turned over. Almost all of the 123 who got the task wrong thought that the *4* card would need to be turned over (Johnson-Laird and Wason 1977).

So what is going wrong here? What sort of reasoning would lead people to think that the *4* card should be turned over? The subjects are asked to identify how they would proceed to disconfirm the hypothesis. So, let *A* be the proposition that the card has a vowel on one side, and *B* the proposition that the card has an even number on one side. Given that the *4* card does have an even number on one side we are given the truth of *B*. Since the subjects think that the card must be turned over they must be reasoning that, given *B* we need to determine whether or not *A* holds in order to evaluate the conditional. One natural way of interpreting this is to think that they must be reasoning along the following lines. If *B* is true and the conditional *If A then B* is true, then *A* must also be true – so we will need to turn over the card to determine whether this is so. But this, of course, is effectively to affirm the consequent. Conversely, subjects ought to be reasoning along the following lines. If *not-B* is true and the conditional *If A then B* is true, then *not-A* must be true – so the *5* card (which effectively gives *not-B*) will have to be turned over to check whether this is so or not. This is a straightforward application of *modus tollens*.

It could be that the experimental subjects, and indeed the rest of us more generally, are reasoning in perfectly domain-general ways, but simply employing the wrong domain-general inferential rules. Instead of applying the domain-general rule of *modus tollens* we all have an unfortunate tendency to apply the equally domain-general, but nonetheless rather unreliable, principle of affirming the consequent. This way of interpreting the results raises all sorts of questions about whether human beings are intrinsically irrational, and so on, but has no implications for how we think about the mechanics of central processing. The idea of central processing as involving domain-general inferential transitions remains intact. It just looks as if quite a few of the inferential transitions that we make are not truth preserving.

However, one of the most interesting aspects of the literature spawned by

the Wason selection task is the powerful evidence it provides that this may well not be the right way to think about the psychology of reasoning. It turns out that performance on the selection task varies drastically according to how the task is formulated.[8] There are "real-world" ways of framing the selection task on which the degree of error is drastically diminished. One striking set of results emerged from a variant of the selection task carried out by Griggs and Cox (1982). They transformed the selection task from what many would describe as a formal test of conditional reasoning to a problem-solving task of a sort familiar to most of the experimental subjects. Griggs and Cox preserved the abstract structure of the selection task, asking subjects which cards would have to be turned over in order to verify a conditional. But the conditional was a conditional about drinking age, rather than about vowels and even numbers. Subjects were asked to evaluate the conditional: If a person is drinking beer, then that person must be over 19 years of age (which is, apparently, the law in the State of Florida). They were presented with the following cards and told that the cards show the names of drinks on one side and ages on the other. Before making their choice subjects were told to imagine that they were police officers checking whether any illegal drinking was going on in a bar.

The correct answers (as in the standard version of the selection task we have already considered) are that the *BEER* card and the *16* card need to be turned over. On this version of the selection task subjects overwhelmingly came up with the correct answers, and relatively few suggested that card *25* would need to be turned over. What is particularly interesting is the subsequent discovery (Pollard and Evans 1987) that if the story about the police officers is omitted, performance reverts to a level comparable to that on the original selection task.

The finding that performance on the selection task can be improved by framing the task in such a way that what is being checked is a condition that has to do with permissions, entitlements and/or prohibitions has proved very robust. This finding is *prima facie* very relevant to the standard view of central processing as involving domain-general rules of inference. At least as far as conditional reasoning is concerned (and there are all sorts of reasons for thinking that conditional reasoning is absolutely fundamental to decision-making and practical reasoning), people seem to be reasoning

8 There is a useful survey of experiments inspired by the original selection task in Chapter 4 of Evans and Over (1996).

according to principles that vary with the subject matter of the inference in question. But then it becomes rather unclear in what sense central processing is domain-general. The fact that we are good at reasoning with so-called *deontic* conditionals (conditionals that express rules, prohibitions, entitlements and agreements) has suggested to many theorists that we have a domain-specific reasoning competence – a competence in a particular type of conditional reasoning that does not carry over to conditional reasoning in other domains.

There are various views about exactly how the relevant class of conditionals is to be demarcated and why there should be such a domain-specific reasoning competence. Perhaps the most influential proposal (and certainly the most controversial) in this area has been that the selection task shows that people are much better at evaluating conditionals involving the detection of cheaters, where a cheater is someone who breaks a rule or who takes a benefit to which they are not entitled. Cosmides and Tooby have worked out this proposal in the context of their overall view of human reasoning and cognitive architecture as *massively modular* (Cosmides and Tooby 1994). They propose that the human mind (perhaps in common with the minds of other higher apes) has a dedicated module for the detection of cheaters – a module that evolved in response to a specific set of problems that confronted our Pleistocene ancestors. This module, the cheater detection module, is just one of a range of highly specialized and domain-specific modules that evolved to deal with specific problems, such as danger avoidance, finding a mate, and so on. According to the massive modularity hypothesis, domain-general reasoning is a myth. What looks like domain-general reasoning is really the operation of domain-specific modules superimposed upon each other by the accidents of our evolutionary history.

But why should there be a cheater detection module? What was the pressing evolutionary need to which the cheater detection module was a response? The massive modularity hypothesis, as its proponents freely admit, is highly speculative – and in many important respects unverifiable. Much of its plausibility rests upon the explanations it offers of why particular modules should have emerged, and the cheater detection module is in many ways a flagship for the program. Cosmides and Tooby's account of the emergence of the cheater detection module is very closely tied to a particular theory of the emergence of cooperative behavior. Biologists, and evolutionary theorists more generally, have long been puzzled by the problem of how cooperative behavior might have emerged. Cooperative behavior presumably has a genetic basis. But how could the genes that code for cooperative behavior ever have become established, if (as seems highly plausible) an individual who takes advantage of cooperators without reciprocating will always do better than one who cooperates? Evolution seems to favor free-riders and exploiters above high-minded altruists. One way of thinking about this problem is by using the model of the prisoner's dilemma discussed in the previous chapter. Many interpersonal interactions (and indeed many inter-

animal interactions) take the form of an indefinitely iterated prisoner's dilemma – that is to say, they involve a series of encounters each of which has the structure of a prisoner's dilemma and where it is not known how many encounters there will be. Let us suppose that the pay-off table for each of these encounters is something like the following.[9]

		Player B	
		DEFECT	COOPERATE
	DEFECT	5, 5	0, 10
Player A			
	COOPERATE	10, 0	2, 2

The pay-off table shows that the dominant strategy for each player is DEFECT.[10] We can think about the problem of the emergence of cooperation in the following terms. Given that the dominant strategy is DEFECT, how can a practice of cooperation ever get sufficiently established to be incorporated in the genotype of the relevant species?

Putting the problem in these terms suggests an answer. As we saw in the previous chapter, social interactions taking the form of indefinitely repeated prisoner's dilemmas can be modeled through simple heuristic strategies in which one bases one's plays not on how one expects others to behave but rather on how they have behaved in the past. The best known of these heuristic strategies is TIT-FOR-TAT, which is composed of the following two rules:

A. Always cooperate in the first encounter
B. In any subsequent encounter do what your opponent did in the previous round

Theorists have found TIT-FOR-TAT a potentially powerful explanatory tool in explaining the evolutionary emergence of altruistic behavior for two reasons.[11] The first is its simplicity. TIT-FOR-TAT does not involve complicated calculations. It merely involves an application of the general and familiar rule that "you should do unto others as they do unto you". The second is that it is what evolutionary game theorists call an evolutionarily

9 We can think of the numbers here as representing abstract units of evolutionary advantage. A prisoner's dilemma emerges from the following two conditions. First, if the first player ranks the possible outcomes in the following order A-B-C-D, then the other player's ranking is D-B-C-A. Player A's most-preferred outcome (where he is the free-rider and player B the sucker) is player B's least-preferred outcome, and vice versa – and each prefers mutual cooperation to mutual defection. The actual numbers are unimportant. The example is a 2-person game, but a similar analysis can be given of a multi-person prisoner's dilemma (the so-called *tragedy of the commons*).

10 See section 7.5 for the reasoning that leads to this conclusion.

11 See Axelrod (1984) and Chapter 12 of Dawkins (1989) for accessible introductions to how TIT-FOR-TAT can be used to explain the emergence of altruistic behavior.

stable strategy – that is to say, a population where there are sufficiently many "players" following the TIT-FOR-TAT strategy with a sufficiently high probability of encountering each other regularly will not be invaded by a sub-population playing another strategy (such as the strategy of always defecting).[12] TIT-FOR-TAT, therefore, combines simplicity with robustness.

Suppose we ask now what needs to be in place for TIT-FOR-TAT to be applied. We saw in the previous chapter that TIT-FOR-TAT does not require any complicated folk psychological machinery. Part of the beauty of TIT-FOR-TAT is that it generates instructions not on the basis of predictions about how other players are likely to behave, or on what one might plausibly take them to believe about the situation, but simply on the basis of how those other players have acted in the past. Nonetheless, simple though TIT-FOR-TAT is, it is not totally trivial to apply. It does not require attributing beliefs and desires and working out how other agents will act in the light of those desires, but it does involve being able to identify instances of cooperation and defection. It involves being able to tell when an agent has taken a benefit without paying the corresponding price (that is to say, when an agent has replied to a cooperative strategy by defecting). Without this basic input, the TIT-FOR-TAT strategy cannot be applied successfully. An agent who consistently misidentifies defectors and free-riders as cooperators (or, for that matter, vice versa) will not flourish. And this, according to evolutionary psychologists such as Cosmides and Tooby, is where the selective pressure came from for the cheater detection module. We evolved a specialized module in order to allow us to navigate social situations that depend crucially upon the ability to identify defectors and free-riders. Since the detection of cheaters and free-riders is essentially a matter of identifying when a conditional obligation has been breached, this explains why we are so much better at deontic versions of the selection task than ordinary versions – and why we are better, more generally, at conditional reasoning about rules, obligations and entitlements than we are at abstract conditional reasoning.

According to the massive modularity hypothesis the proposed cheater detection module is a model for understanding the mind as a whole. To use a popular analogy, the mind is a Swiss Army knife, with a wide range of specialized tools, each designed to carry out a different task. There is no such thing as central processing and there are no domain-general principles of reasoning. All reasoning involves domain-specific principles tailored to particular types of subject matter. There is no sharp distinction to be drawn either at the personal level or at the level of cognitive architecture between central processing and peripheral processing. The entire mind is modular.

So, how seriously should we take the massive modularity hypothesis? The

12 For further discussion see Chapter 3 of Skryms (1996).

hypothesis has come under attack from partisans of more traditional approaches to the mind. Jerry Fodor in particular, has argued that the massive modularity hypothesis is incoherent (Fodor 2000). There could not, he argues, be a cognitive system that is completely modular. The problem emerges when we think about the relation between the *Darwinian modules* proposed by supporters of the massive modularity hypothesis and the more familiar types of modules that have been proposed in models of early visual processing and other forms of peripheral processing (what we might term *Fodorean modules*, in deference to Fodor who first proposed them). The central feature of any modular system, whether it is Darwinian or Fodorean, is that it takes only a limited range of inputs. So, the obvious question anyone proposing a modular cognitive capacity has to answer is how that limited range of inputs is selected. In particular, they need to specify whether any processing is involved in identifying the relevant inputs and discriminating them from inputs that are not relevant.

For classical Fodorean modules the answer is straightforward. Modules responsible for low-level tasks such as early visual processing and syntactic parsing are supposed to operate directly on sensory inputs and it is usual to postulate sensory systems (so-called *transducers*) that directly filter the relevant inputs. These filters ensure, for example, that only information about light intensity feeds into the earliest stages of visual processing. One obvious difference, however, between Darwinian modules and Fodorian modules is that they operate on fundamentally different types of input. The inputs into the cheater detection module, for example, must be representations of social exchanges of the sort that may be exploited by cheaters. Indeed, if we take the experimental inspiration for the cheater detection module seriously, the inputs must be represented in a form somewhat akin to a deontic conditional. It seems clear, therefore, that some processing is required to generate the appropriate inputs for the cheater detection module. It does not make sense to postulate the existence of social exchange transducers. There has to be some sort of filtering operation that will discriminate all and only the social exchanges – and indeed this filtering will have to be sophisticated enough to identify appropriate versions of the Wason selection task as instances of social exchanges.

This is where Fodor's objection strikes. According to the massive modularity hypothesis, the processing involved in this initial filtering must be modular. Clearly, the filtering process will only work if the filtering module has a broader range of inputs than the module for which it is doing the filtering. But, on the other hand, since the filtering process is modular, it must have a limited range of inputs. The filtering process is itself domain-specific, working to discriminate the social exchanges from a slightly broader class of inputs – perhaps a set of inputs whose members have in common the fact that they all involve more than one person. So the same question arises again. How is this set of inputs generated? Presumably a further set of processing will be required. *Ex hypothesi,* this processing will itself be modular.

So once again a further set of domain-specific inputs needs to be identified. Eventually, Fodor argues, we will end up with processing that is so domain-general that it can hardly be described as modular at all. A similar line of argument will apply to all the other Darwinian modules. The massive modularity hypothesis collapses, because it turns out that massive modularity requires complete domain-generality.

The argument here is characteristically ingenious. It is not clear, however, that it really establishes what Fodor is trying to establish. Suppose we grant its conclusion, namely, that each Darwinian module presupposes a lengthy process of filtering that will have to begin with processing that is pretty much domain-general. It clearly follows from this that there will have to be *some* kind of domain-general processing. What is not clear, however, is that the domain-general processing involved will be of the type that the massive modularity hypothesis is committed to ruling out. The force of the massive modularity hypothesis, as we have seen, lies in its denial that there are domain-general processes of decision-making and belief fixation – and, in particular, in its denial that there are domain-general principles of inference that hold across the holistic system of propositional attitudes. But this claim does not seem to be affected by Fodor's argument. Fodor may have established that the massive modularity hypothesis requires domain-general processing, but not that it requires domain-general reasoning. What he would need to show is that these ever more domain-general filtering modules can only determine the relevant inputs for the Darwinian modules in ways that require sensitivity to the global properties of belief systems. But why should fixing whether an event is a social exchange, for example, require such *global* sensitivity? Why should it involve bringing to bear one's background knowledge? Why should it involve domain-general inferential principles? The filtering process seems far more akin to high-level perceptual processing than to central processing as traditionally construed. To return to an earlier theme, fixing the inputs for Darwinian modules may well best be understood as a form of template-matching and relatively straightforward pattern recognition.

But although Fodor's objection in principle to the massive modularity hypothesis fails to carry conviction, the opponent of massive modularity is not without resources. It seems highly likely that, for any Darwinian module, there will be some potential inputs that fall clearly within its domain. There will be other potential inputs that are only borderline candidates for that particular module. So, for example, to remain with the cheater detection module, some social exchanges will clearly be the sort of situation where the detection of free-riders is significant, and the experiments that we have been considering are designed to tap into situations of this type. These would include situations where explicit prohibitions/permissions are directly salient. The drinking age examples fall into this category. So too do other staples of the experimental literature, such as the immigration officer checking whether passports are suitably stamped with evidence of the required inoculations. Also clear-cut are situations where a

benefit is being received in exchange for a reciprocal benefit – and hence where there is a threat of free-riding. For many social situations, however, it is far from clear whether to classify them as social exchanges. There may be other modules to which they are equally relevant. Something might be a social exchange when looked at from one point of view, but a potentially dangerous situation when looked at from another. Let us call this situation *S*. Under the first description *S* would be an input for the cheater detection module, while under the second description *S* might be relevant to the danger avoidance module. Fodor's argument is effectively that considerable domain-general processing is required in these sorts of situation in order to determine which module should come into play – in order to determine *S*'s point of entry into the cognitive system. In many ways, however, it seems more plausible that the representation of *S* will be recruited by any system to which it is potentially relevant, so that a given representation of a single situation may well be processed in parallel by a range of different Darwinian modules.

This seems to follow from the response already given to Fodor. If the process of filtering inputs for each domain were indeed perceptual and based on template matching, then one would expect it to be relatively coarse-grained. This seems even more likely when one takes into account the computational costs of filtering mechanisms that are completely accurate. Recall that we are taking seriously the hypothesis that Darwinian modules evolved in response to specific problems faced by our pre-hominid ancestors. We have to bear in mind that evolution is a "satisficer", and is correspondingly unlikely to have produced classificatory mechanisms that only ever come up with the right answer. Our susceptibility to optical and other sensory illusions is good evidence of that. So, the question is what errors the system is likely to make. One might think that the costs of having a system that produces a number of false positives (that is to say, a system that wrongly classifies social situations as social exchanges when they are not) is more likely to have evolved than one that produces a number of false negatives (that is, a system that fails to classify social exchanges as social exchanges). This is even more plausible in the case of predator and danger avoidance modules, where the costs of even a single false negative can be terminal.

If this is right, then it seems likely that a significant number of representations will be processed in parallel by more than one Darwinian module. This will create a processing problem. The outputs of the relevant module will need to be reconciled if, for example, the predator avoidance module "recommends" one course of action and the cheater detection module another. The cognitive system will have to come to a stable view, prioritizing one output over the other. Clearly, therefore, further processing will be required. And it seems likely that this will be domain-general processing in the sense that we have been discussing. It will involve bringing background knowledge to bear in order to determine which course of action is required. The principles of reasoning deployed in making this decision cannot be domain-specific.

Intramodular principles of reasoning cannot be used to resolve conflicts between modules. The principles of reasoning need to be applicable to both of the relevant domains, and indeed to any other domains that might be potentially relevant. It seems very plausible that this type of processing of the *outputs* of Darwinian modules will have to involve precisely the sort of sensitivity to the global properties of the propositional attitude system that does not seem required to filter and process the *inputs* to those modules.

The general thought here is really rather straightforward. According to the massive modularity hypothesis the mind is a complex structure of superimposed Darwinian modules that have evolved at different times to deal with different problems. Given the complexities of human existence and human social interactions, there will have to be a considerable number of such modules. Given those very same complexities, moreover, it seems highly unlikely that every situation to which the organism needs to react will map cleanly onto one and only one Darwinian module. It is far more likely that in many situations a range of modules will be brought to bear. Something far closer to what is standardly understood as central processing will be required to reconcile conflicting outputs from those Darwinian modules. This central processing will be *un*encapsulated and domain-general.

If this line of reasoning is correct, then the massive modularity hypothesis ought to be treated with some suspicion as a *complete* account of human cognition and human cognitive architecture (although it may well be true of organisms that respond in less finely-tuned ways to their environment). Nonetheless, there are important lessons to be learned from it. If there are indeed such things as Darwinian modules (and this of course is ultimately an empirical matter), then this means that the standard view of the route from perception to action needs to be substantially modified. Although, for the reasons I have suggested, there is still a significant role for domain-general central processing to play, this role is much less significant than it is on the standard view. Whereas, on the standard view, central processing is required for all behavior that is not purely reflex or controlled by something like an innate releasing mechanism, the massive modularity hypothesis draws our attention to the possibility that various types of sophisticated behavior may well bypass central processing entirely. The massive modularity hypothesis opens up the possibility of more or less direct links between perception and action that are sophisticated enough to be characterized as forms of intentional behavior, and yet that do not engage the propositional attitude system.

In this sense, therefore, the massive modularity hypothesis, together with the other criticisms of the standard view of the route from perception to action that we have looked at, fits very well with some of the suggestions that emerged in the previous chapter about the scope of commonsense psychology. The proposal there was that commonsense psychology might play a far less prominent role in social understanding and social coordination than

almost all philosophers and most cognitive scientists believe. It may well be, I suggested, that we deploy a range of mechanisms and heuristics for understanding other people and navigating social interactions, mechanisms and heuristics that do not involve reasoning about the beliefs and desires of other agents. We do of course use commonsense psychology in the service of explanation and prediction – but perhaps far less frequently than is commonly assumed. What has emerged in this chapter is that propositional attitudes may play a correspondingly smaller role in generating action. Whereas the standard view of the route from perception to action holds that domain-general reasoning involving the propositional attitude system is involved in all but the simplest and most straightforward types of behavior, the picture of cognition and cognitive architecture that is emerging downplays the significance of domain-general processes of belief fixation and decision-making. Once again, it is not that they never come into play. As we have seen, it seems unlikely that the massive modularity hypothesis can be a complete account of cognition and cognitive architecture. Nonetheless, they may well come into play far less frequently than is generally thought.

9 Propositional attitudes

Contents and vehicles

- Another look at the interface problem
- The argument for structure
- The problem of structure in artificial neural networks
- Rejecting the structure requirement
- Finding structure in artificial neural networks
- Overview

The previous two chapters have concentrated on the nature and scope of commonsense psychology and on very broad questions to do with the architecture of cognition. We looked in Chapter 7 at the role of propositional attitude psychology in social understanding and coordination, while Chapter 8 explored the idea that there is a fundamental distinction at the level of cognitive architecture between, on the one hand, peripheral processes that are modular and do not involve the propositional attitudes and, on the other, non-modular central processes defined over the propositional attitudes. The overarching theme of both chapters was the role of the propositional attitudes in cognition. In this chapter we will place these general questions to one side and instead concentrate on more local issues concerning the propositional attitudes – local issues that arise whether one thinks that the scope of commonsense psychology is broad or narrow, and whether or not one thinks that the modular/non-modular distinction can be maintained at the level of cognitive architecture.

Almost all the views we are considering, with the possible exception of some extreme proponents of the neurocomputational approach to the mind, are agreed that the propositional attitudes have *some* role to play both in explaining behavior and in generating behavior. This minimal commitment is all that is required to set up the problems that will be discussed in this chapter. The principal issue we will be tackling is the question of how propositional attitudes must be realized in the nervous system in order for us to be able to appeal to them in explaining behavior. Do the ways we use propositional attitudes in explaining cognition and behavior place any constraints upon how we think about the *vehicles* of those attitudes?

To make progress on this issue we need to return more directly to the interface problem of explaining how commonsense psychological explanations connect up with levels of explanation lower in the explanatory hierarchy. Section 9.1 draws together some of the strands of the discussion in the previous chapters and explains their potential implications for how we think about the interface problem. Section 9.2 outlines the argument put

forward by language of thought theorists for the thesis that propositional attitudes must have structured vehicles – that is, for the thesis that propositional attitudes must be physically realized in a form that shares the structure of the content of the attitude. In section 9.3 we will see why artificial neural networks do not, *prima facie*, seem to allow for appropriately structured vehicles. In section 9.4 we will look at two ways of reacting to this *prima facie* tension between the *structure requirement* and artificial neural networks. The tension might either be exploited as a direct argument for the language of thought hypothesis, or it might form the basis of an eliminativist argument to the effect that we are simply mistaken in appealing to propositional attitudes in psychological explanation. In section 9.5, we consider whether the tension between the structure requirement and the neurocomputational approach to the mind is as clear-cut as it initially seems. We will look at proposals for identifying structure in artificial networks, and explore whether the structure requirement should be imposed as strictly as it is in the arguments considered in section 9.4.

9.1 Another look at the interface problem

In Chapter 2 I characterized the interface problem as follows:

> *The interface problem* How does commonsense psychological explanation interface with the subpersonal explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy?

The four pictures of the mind that we have been considering offer very different approaches to the interface problem. The picture of the autonomous mind sets out to *deflate* the problem, arguing that there is a radical incommensurability between norm-governed commonsense psychological explanation and the various subpersonal levels of explanation. But the other three pictures take a more positive approach. The pictures of the functional and representational mind take a relatively fixed view of the nature of commonsense psychology, proposing to proceed in a top-down way by looking for the subpersonal *vehicles* of commonsense psychology. Proponents of the neurocomputational approach, in contrast, offer a *co-evolutionary research program* that is simultaneously top-down and bottom-up, with our understanding of commonsense psychology co-evolving with our understanding of the neural basis of cognition.

The picture of the autonomous mind will not feature in this chapter. The problem that we will be dealing with emerges from assumptions that supporters of the autonomous approach are unlikely to accept. Autonomy theorists can only accept that propositional attitudes have subpersonal vehicles in a very limited sense. As we saw in Chapters 3 and 6, some autonomy theorists

(with Dennett as the most prominent example) think of personal-level psychological states as emergent in a way that rules out the possibility of theorizing about their subpersonal vehicles. Other autonomy theorists (such as Davidson) do accept that personal-level psychological states can be realized in physical structures at the subpersonal level. However, the account that Davidson offers of the vehicles of propositional attitude states is highly circumscribed. Mental events are token-identical with physical events, where a physical event is one that can be individuated and characterized in terms of the language of physics. Davidson only offers us these two ways of thinking about propositional attitudes. We can think of them either as mental events governed by the norms of rationality and featuring in forms of explanation that are irreducible to other forms of explanation, or as events under physical descriptions that feature in physical laws. What we cannot do, however, is think of them as cognitive states at the subpersonal level. If we think of them as cognitive then we have to think of them at the level of propositional attitude psychology. If we do not think of them as propositional attitude states, then we have to think of them as non-cognitive. There is no possibility of an interface between personal-level psychology and the various subpersonal levels of psychological explanation lower down in the explanatory hierarchy.

Nor will psychological functionalism as an approach to the mind have a large role to play in this chapter. This is because this chapter will focus primarily on how we should understand the vehicles of *individual* propositional attitudes, an issue that tends not to be at the forefront of discussions of psychological functionalism. One of the things that make psychological functionalism distinctive is that it focuses primarily on the explanation of mechanisms, and in particular on using the methodology of functional decomposition to show how complex tasks can be broken down into simpler tasks that can be carried out by simpler mechanisms. This general approach has significant implications at the level of cognitive architecture. However, since the focus of psychological functionalism is upon tasks and the mechanisms that perform them, it places correspondingly less emphasis on individual propositional attitudes, which do not bear the principal explanatory weight of the approach in the way that they do, say, in philosophical functionalism or the picture of the representational mind. We will, accordingly, be concentrating on these latter two approaches to the mind and to psychological explanation in this chapter.

In the last two chapters we have been looking at the nature and scope of commonsense psychological explanation. The way these are understood is significant for how the interface problem is to be addressed. There are ways of understanding the nature and scope of commonsense psychological explanation that clearly count in favor of the representational and functional pictures of the mind. Both philosophical functionalism and the representational picture are firmly rooted in a conception of commonsense psychological explanation as primarily propositional attitude explanation. Theorists from

both approaches tend to be committed to the idea that our fundamental way of making sense of other people and of each other is through interpreting their behavior in the conceptual framework of commonsense psychology, and hence to what in Chapter 7 we discussed as a broad conception of the domain of propositional attitude psychology. Philosophical functionalists and representational theorists tend also to be committed to what in Chapter 8 I characterized as the standard view of the route from perception to action, which draws a sharp distinction to be drawn at the level of cognitive architecture between non-modular central processing, which is the domain of the propositional attitudes, and modular peripheral processing, which encompasses the inputs to and outputs from the propositional attitude system.

It has emerged from the discussion in the previous two chapters, however, that commonsense psychological explanation may not be quite as straightforward as it is often taken to be. Philosophical functionalists and representational theorists identify commonsense psychological explanation and propositional attitude psychology. Yet, as we saw in Chapter 7, there is an important ambiguity in the notion of commonsense psychology. At the most general level we can think about commonsense psychology as the complex of skills and abilities that allows us to succeed in understanding other people and in coordinating our behavior with theirs. When commonsense psychological explanation is understood in this sense, it is very much an open question whether those skills and abilities actually involve applying the concepts and categories of propositional attitude psychology. There could well be a wide range of skills and abilities underpinning social understanding and social coordination that do not involve exploiting the machinery of propositional attitude psychology. The candidates we considered included heuristics such as TIT-FOR-TAT, social scripts and routines, and mechanisms that detect and respond to other people's emotional states. If we expand our notion of commonsense psychological explanation to include these additional factors, then the interface problem becomes more complicated. It may start to seem unlikely, for example, that there will be a single way of responding to the interface problem. Commonsense psychology has different aspects and one might expect different answers to the interface problem depending upon which aspect of commonsense psychology is in play. One possible way of thinking about commonsense psychology would be in terms of a core of propositional attitude psychology, surrounded by a more extensive periphery of heuristics, template-matching mechanisms, scripts, routines, and so forth. Solutions to the interface problem would look very different depending on whether one was focusing on the core or on the periphery. It might turn out, for example, that whereas some version of philosophical functionalism or the representational picture of the mind is more appropriate for the core of commonsense psychology, the neurocomputational approach offers a better way of thinking about the periphery.

One suggestion that emerged in Chapter 8 is that many of the tools we employ for social understanding and social coordination might turn out to

be complex forms of pattern recognition and template-matching. Mechanisms for detecting and responding to other people's emotional states seem to fall clearly into this category. So too do the ways of interpreting other people's behavior involved in applying heuristics such as TIT-FOR-TAT. The basic rule of TIT-FOR-TAT is to begin by cooperating and then to copy the behavior of the other participants in social interactions – to cooperate when they cooperate and to defect when they defect. This requires being able correctly to identify when people have cooperated and when they have defected. It is plausible to think that this is a form of template matching. The same holds of the social scripts and routines that govern many of our social interactions. The fact that template-matching has so significant a role to play in social understanding and social coordination is very important for the interface problem because of the demonstrable success of artificial neural networks in modeling processes of template-matching (Bechtel and Abrahamsen 1991, Chapter 4). Although there has been little attempt to construct networks that can solve problems of social understanding and social coordination, the type of task that such networks would have to perform seems very similar to the type of task on which the performance of artificial neural networks has been well studied. There are many connectionist models of face recognition, for example (McLeod *et al.* 1998, Chapter 13). One would expect similar sorts of models to be able to identify and categorize emotions. Reading emotions from facial expression is the sort of associative learning task at which connectionist networks excel.

This seems an area where a co-evolutionary research program might flourish. We do not yet have a very clear or determinate conception of how to explain social understanding and social coordination at the personal level – or rather, the further we move from the idea that all social understanding and social coordination is to be explained in terms of propositional attitude psychology the less clear and determinate our personal-level understanding becomes. It seems likely that the outcomes of attempts to model social understanding and social coordination will feed directly into our personal-level accounts of those phenomena. Discovering, for example, that a particular social task can be successfully carried out by an artificial neural network might be a powerful reason for thinking that it is best viewed at the personal level as a template-matching task. Similarly, placing pressure at the personal level on the broad conception of the scope of commonsense psychology might lead experimenters to attempt to provide connectionist models of tasks of social understanding and social coordination.

There is room, therefore, for a two-layer response to the interface problem, a response that draws on the resources of more than one of the pictures of the mind we have been considering. The two-layer response might combine a neurocomputational approach to the peripheral aspects of commonsense psychology with a representational or functional approach to the propositional attitude core. This idea of a two-layer response fits well with some of the points about large-scale cognitive architecture that emerged in

Chapter 8. We saw there how pressure can be placed upon the standard view of cognitive architecture as involving peripheral modules that process input to and output from a central processing system deploying the propositional attitudes. There may be relatively sophisticated connections between perception and action that involve relatively sophisticated forms of processing and yet that completely bypass central processing. This is one of the lessons to be drawn from the massive modularity hypothesis (however suspicious one is of the massive modularity hypothesis as a complete account of cognition). These connections, which might be Darwinian modules or specialized neural circuits, could be models for the subpersonal mechanisms subserving the periphery of commonsense psychology – and indeed, in the case of mechanisms such as the cheater detection mechanisms, they may well actually be the relevant subpersonal mechanisms. In any case, it is at least possible that the cognitive architecture required to support these mechanisms is fundamentally different from the cognitive architecture required by the propositional attitude system.

We will be returning to the two-layer response and to the idea that solving the interface problem might require appealing to more than one type of cognitive architecture. We will approach it by examining what might seem a more circumscribed issue relevant only to the propositional attitude component of commonsense psychology. This is the issue of whether the demands of psychological explanation impose certain constraints upon the subpersonal vehicles of propositional attitudes. Proponents of the language of thought hypothesis have argued that we cannot make sense of the causal dimension of propositional attitude psychology unless we take the vehicles of beliefs, desires and other attitudes to have a structure isomorphic to the structure of the content of the relevant attitude. This requirement of isomorphism can only be satisfied, it is argued, if the vehicles of propositional attitudes are sentences in an internal language of thought. There is a lively debate both about whether artificial neural networks can support representations that have the requisite structure, and about the legitimacy of the demand for structure. The debate is standardly pursued on the assumption that there is a single cognitive architecture and hence that there is a straightforward choice between the artificial neural networks approach and the language of thought hypothesis. However, as will become clearer in section 9.3, the arguments and counter-arguments in the debate leave open the possibility of something like the two-layer response.

9.2 The argument for structure

Whatever positions are taken on the scope of commonsense psychology and the issue of how best to think about the overall architecture of the mind, almost all parties to these debates accept, first, that there are propositional attitudes; second, that the propositional attitudes are in some sense realized at the subpersonal level; and, third, that propositional attitudes have a role

to play in explaining behavior. The various different approaches differ on how significant a role the propositional attitudes play in explaining behavior, and on how exactly the propositional attitude system is subpersonally realized. In this section we will be looking at an influential way of thinking about how propositional attitudes are, or could be, realized at the subpersonal level.

It will be helpful to recap some basic points about propositional attitudes. A propositional attitude, as standardly construed, should be understood in terms of two components – a content, and an attitude taken towards that content. If I am correctly described as having the belief that *p*, for example, this is standardly taken to mean that there is a content (the content that *p*) to which I take the attitude of belief. One rationale for distinguishing content and attitude is that different people, and indeed the same person, can take different attitudes to the same content. I can believe to be the case, for example, what you hope to be the case. And I can come to desire to be the case what I formerly feared to be the case. Some philosophers use the term 'force' to capture the "attitudinal" component of a propositional attitude. The force of a propositional attitude can be understood in psychological terms. Different attitudes play different causal roles within the cognitive economy. Beliefs and desires interact differently with other propositional attitudes and have fundamentally different implications for behavior.

The content of an attitude, in contrast, is an abstract entity. There are many different ways of thinking about what this abstract entity might be. It might be what is sometimes known as a Russellian proposition – that is to say, a structured complex of individuals and properties. Alternatively, it might be a Fregean proposition made up of individual senses (or concepts). For present purposes it does not matter which, if either, of these two accounts is given. From the viewpoint of the philosophy of psychology we are interested primarily in the vehicles of propositional attitudes. As we have had frequent occasion to stress, it is widely held that propositional attitudes interact causally with each other and with other mental states. It is clear, however, that these causal interactions cannot take place at the level of the content of those attitudes. Abstract entities cannot interact causally with each other, however they are understood. It is no more (and no less) easy to understand how two structured complexes of individuals and properties can interact with each other than it is to understand how two Fregean propositions can interact causally with each other.

Many theorists have concluded, therefore, that we can only understand causal interactions between propositional attitudes if we think of the contents of those propositional attitudes as realized in physical structures in the mind/brain. These physical structures are the vehicles of the propositions in question. This inference is rejected by an influential minority of theorists, whose views have already been considered in Chapter 6. We considered there two different ways of arguing that propositional attitude explanation can be

a species of causal explanation without assuming the existence of causally efficacious internal items. The first was Dennett's suggestion that commonsense psychological explanations track genuinely existing patterns in the behavior of organisms (and cognitive systems more generally), but not in a way that requires the existence of independently identifiable and causally interacting physical structures. The second was the counterfactual approach, according to which the truth of causal statements about cognitive systems lies simply in the truth of certain counterfactual statements about how the system in question would have behaved in different situations. There is no need to rehearse the arguments for and against these views once more. This chapter will proceed on the assumption that commonsense psychological explanation can only be a species of causal explanation if there are causally efficacious inner items that serve as the vehicles of the propositional attitudes cited in those explanations. Those who are not convinced of this assumption may nonetheless be interested in investigating its consequences.

We need both to give an account of how propositional attitudes can have causally efficacious inner items, and to explain how propositional attitudes can be causally efficacious in virtue of their contents. Propositional attitude explanations work on the assumption, not simply that beliefs, desires and other propositional attitudes cause behavior, but that they cause behavior in virtue of how they represent the world. We might say that propositional attitudes have two dimensions (a causal dimension and a representational dimension) that must be kept in harmony. But how are we to secure this harmony? If, as we have decided to accept for the sake of argument, the causal dimension of commonsense psychological explanation needs to be understood in terms of the physical structures that realize those attitudes, then it is natural to think that, if the harmony is secured, it can only be in virtue of certain features of those physical structures. Since the causal dimension of propositional attitude explanation is determined by the physical vehicles, then it looks very much as if all those features of the content of the attitude that are not reflected in the physical vehicle will drop out of the picture.

This line of argument has been pressed most forcefully by proponents of the language of thought hypothesis, who highlight one feature that (they maintain) must be possessed by the vehicles of propositional attitudes if propositional attitude explanation is to count as genuine causal explanation. Language of thought theorists hold that the causal dimension of propositional attitude explanation stands or falls with the vehicle of a given propositional attitude having a structure isomorphic to the structure of the content of that attitude. To say that one structure is isomorphic to another is to say that each element in one has a corresponding element in the other, and neither structure has an element to which nothing corresponds in the other structure. If the elements in each structure are related to each other in analogous ways, then the two structures can be put into what mathematicians and logicians call a one–one correspondence.

Consider, for example, the belief that Chicago is north of St Louis. If we understand the content of belief in Russellian terms, then the content of this belief is given by a Russellian proposition composed of two individuals and a relation holding between those two individuals. The structure of this Russellian proposition would be represented in the predicate calculus by the formula '*aRb*' where '*a*' and '*b*' pick out Chicago and St Louis and '– *R* –' stands for the relation of being to the north of. According to the language of thought theory, the vehicle of this belief (the physical structure realizing it in the central nervous system) must have an isomorphic structure. It must be made up of three distinguishable and separable elements, each of which corresponds to one element in the Russellian proposition. Just as the content of the belief is a complex entity made up of Chicago, St Louis and the relation of one thing being to the north of another, the vehicle of that belief is a complex physical structure made up of physical components standing in for Chicago, St Louis and the relation held to hold between them. To say this, moreover, is effectively to say that the vehicle of the belief has the structure of a natural language sentence (on the very plausible assumption that two things that can be put into a one–one correspondence have the same structure – after all, what other notion of sameness of structure do we have?). This sentence is of course the sentence that expresses the belief in question. The central claim of the language of thought hypothesis, then, is that the vehicles of propositional attitude contents are physical structures with separable and recombinable components that can be put into a one–one correspondence with the structure of the sentence that expresses the content of the belief.

Why is the causal efficacy of propositional attitude explanation supposed to depend upon propositional attitudes being realized in physical structures that are isomorphic to the contents of those attitudes? The basis thought is that propositional attitude explanation can only be applied to cognitive systems capable of certain forms of thought. Propositional attitude explanation assumes, for example, that the organism whose behavior is being explained is sensitive to the logical consequences of its beliefs. This does not mean, of course, that the organism should believe all the consequences of everything it believes, but simply that it should be able to move beyond its current beliefs in line with some of the logical implications of what it believes. Similarly, the organism must be able to see connections between its propositional attitudes, to bring its beliefs and desires into harmony. A good way of thinking about what drives the language of thought theory is through the idea that these very general constraints upon what it is to be a thinker (and hence what it is to have one's behavior usefully explained and predicted in psychological terms) impose certain requirements upon the vehicles of propositional attitudes.

So, what are these general constraints upon what it is to be a thinker? And how exactly do they impose further requirements upon the vehicles of propositional attitudes? Proponents of the language of thought hypothesis

stress two very general aspects of thought – aspects that are, they claim, so essential to thought that no cognitive system that lacked them would be able to count as a thinker. These aspects are best understood as types of ability that a thinking system must have.

The first ability is the ability to generate and understand indefinitely many new thoughts. Thinkers may be limited in the number and type of thoughts they can think by considerations of time, energy, and so on, but the nature of thought itself imposes no such constraints. There is an analogy with language mastery that it is worth bringing out. What distinguishes genuine understanding of a language from the type of disjointed ability to communicate that comes from parrot-style learning of a few sentences from a phrase-book is that the genuine language-user can combine words to form new sentences and is capable of understanding novel combinations of words that she has not previously encountered, whereas the phrase-book user is confined to a fixed repertoire of set phrases. Thought, it seems, has the same characteristic of *productivity* or *generativity*. Indeed, the productivity of thought seems closely linked to the productivity of language. Given that language essentially communicates thoughts, no language-user could produce and understand new sentences unless they were capable of productive thought. This type of generativity is implicated, moreover, in key cognitive abilities presupposed by the practice of psychological explanation – such as the ability, for example, to extend one's beliefs in line with their logical commitments.

The second key cognitive ability for language of thought theorists is closely related to the first. In fact, it can be viewed as a fundamental way of achieving the productivity of thought. Grasping a new thought is not like learning a new phrase in a phrase book – any more than generating a new sentence is. Just as sentences are made up of words and understood in terms of the meanings of the words that make them up, so too are thoughts grasped in terms of the individual constituents that make them up. These individual constituents are concepts, on a broadly Fregean view, or individuals and properties, on a broadly Russellian view. Grasping a novel thought is rather like constructing a novel sentence. It can be a matter of putting familiar things together in new ways. Alternatively, it can be a matter of putting new things together in familiar ways. Either way, however, it depends upon there not being any fixed rules about what can combine with what. In language there are, of course, rules of grammar that prevent us, for example, from putting nouns where verbs ought to be, or from using adverbs to qualify pronouns. But there are no rules about which nouns can go together with which verbs, or which adverbs can qualify which adjectives. Within the very general constraints imposed by the laws of grammar the possibilities of combination are unlimited (although of course some combinations will be more meaningful than others). Exactly the same holds for thoughts. The elements of which thoughts are made up can be combined in any way permitted by the general laws of logic. This feature of

language and thought is generally termed *systematicity*. It is easy to see once again how systematicity is closely linked to the types of inferential ability that are presupposed by practices of psychological explanation.

Suppose we accept, then, that any system of genuine thought must be productive and systematic. What constraints does this impose upon the vehicles of those thoughts? The key claim of language of thought theorists is that the systematicity and productivity of thoughts require the vehicles of those thoughts to be composed of separable and recombinable physical elements that map individually onto the distinct elements of the contents of those thoughts. Because the content of any given propositional attitude is an abstract object, the structure of that content cannot be directly exploited in cognition. Cognition, we are assuming, is in the last analysis a physical process – and indeed a causal process. There must therefore be a physical surrogate that allows the structure of the content to be exploited in thought and reasoning. This physical surrogate (which is, of course, the vehicle of the content) must have the same structure as the content if it is to allow that structure to be exploited. Since the structure of a content is given by the logical structure of the sentence that expresses that content, it follows that the physical structure must itself be isomorphic to the logical form of that sentence. And that is all that is meant by saying that the vehicle of a propositional attitude is a sentence in the language of thought.

The picture of the representational mind argues from certain very general requirements upon what it is to be a thinker that the vehicles of propositional attitudes must be sentences in an internal language of thought – or, alternatively, that propositional attitude contents must be realized by physical structures that can be put into a one–one correspondence with the logical form of the sentence that expresses the relevant content. The argument raises a range of questions. One might wonder about the putative requirements of systematicity and productivity. Are they really as non-negotiable as they are taken to be by the language of thought theorist? Or one might wonder about the argument from systematicity and productivity to structure at the level of vehicle. Is it so obvious that systematicity and productivity could not emerge from structured contents with unstructured vehicles? These and related questions set the agenda for the remainder of the chapter.

9.3 The problem of structure in artificial neural networks

Chapter 5 considered the neurocomputational approach to the mind. This approach is driven by the thought that the mind cannot be studied in a purely top-down or bottom-up manner. We will make progress only by allowing a two-way influence between the personal and the subpersonal levels of explanation – and in particular by looking closely at the relation between models of neural functioning and our personal-level psychological

concepts and modes of explanation. This *co-evolutionary research methodology* clearly leaves open the possibility that our personal-level psychological concepts and modes of explanation may have to be revised in the light of our best models of neural functioning. This line of reasoning has been taken to the extreme by *eliminativist* supporters of the neurocomputational approach, who argue that our models of neural functioning *falsify* commonsense psychological ways of thinking about the mind (Churchland 1986; Stich 1983). But the neurocomputational approach leaves open the possibility of revision without rejection. It is of the essence of the co-evolutionary approach that the influence can work in both directions, and it may well be that we may need to rethink our models of neural functioning in the light of constraints imposed by our theories of personal-level psychology.

An interesting dialectic in this area is created by what seems to be a basic tension between our current best models of neural functioning and the requirements upon the vehicles of propositional attitudes sketched out in the previous section. In brief, if we model neural functioning with artificial neural networks, and if we assume that artificial neural networks are good guides to the broad features of the vehicles of propositional attitudes, then it is difficult to see how the vehicles of propositional attitudes can be structured in the manner characterized in the previous section.

Let us begin by looking at the *prima facie* tension between artificial neural networks and the (putative) requirement of structure. If the arguments of section 9.2 are sound, then any given propositional attitude must have a structured vehicle that can be mapped onto the sentence that gives the logical form of the content of that attitude. We can think about this mapping in terms of two requirements. First, the vehicle of the attitude must be composed of separable elements that correspond to each of the relevant components of the propositional attitude content. Second, those elements must be common to a range of propositional attitude vehicles. Suppose I have the beliefs that I would express with the following three sentences "St Louis is north of New Orleans", "New Orleans is in Louisiana" and "Louisiana borders on the Gulf of Mexico". According to the view we are considering, the vehicles of these beliefs will have elements that are common to more than one vehicle. There will be an element corresponding to the proper name "New Orleans" in the vehicle of the first and second belief, and an element corresponding to the proper name "Louisiana" in the vehicles of the second and third beliefs – just as the sentences that express those beliefs have common elements.

It is difficult to apply this way of thinking to artificial neural networks. We can approach this by thinking about the structure of a sentence (since this is our principal way of thinking about the structure of a propositional attitude content, or the structure of the vehicle of that content). An integral part of what it is for a sentence to be structured is that different parts of the sentence should do different jobs – that there should be a verb, for example, that is clearly distinct from the subject or object of that verb. If a sentence

has an element that does one job, then that element will do the same job in another sentence in which it appears (cases of ambiguity apart). The different parts of a sentence tend to do different things, and moving them around will frequently change the meaning of the sentence. The same holds, *pari passu*, for the different elements in the content of a propositional attitude. In contrast, perhaps the most striking feature of neural networks is that they are characterized by their homogeneity. They are composed of levels of units that behave in very similar ways. What a unit in an artificial neural network does is vary its activation level as a function of the levels of activation transmitted to it by the input units to which it is connected (or, if it is an input unit, by the appropriate activation from outside the network). The activation level of each unit is then passed on to the units to which it is connected in the next layer – and so on until the output layer is reached. The most important feature of an artificial neural network is the pattern of weighted connections holding between individual units. The individual units are relatively unimportant. If the weighted connections were held constant, then exchanging any two units in the network would leave the behavior of the network completely unchanged.

If we think of computation in the broadest terms as the process by which an input into a given cognitive system is transformed into an output, then computation in an artificial neural network takes the form of spreading patterns of activation over the layers intervening between input and output. A network learns by adjusting its weights in accordance with a particular learning algorithm and in response to feedback from outside the network – feedback that conveys information about the degree of discrepancy between the actual output and the desired output. Both of these processes seem very distant from the type of transformation of input into output that the language of thought hypothesis is intended to capture.

For proponents of the representational picture of the mind, computation is essentially a matter of the manipulation of symbol structures in ways that are sensitive only to formal properties of the symbols in question. The representational approach depends upon computation being defined over symbol structures that permit a sharp distinction between syntax and semantics, in the following two senses. First, it must be possible to consider and manipulate these symbol structures in complete abstraction from their representational properties, just as one can manipulate a well-formed formula of the predicate calculus without having any idea of what the elements of that formula stand for (and hence of how the formula as a whole should be interpreted). Second, it must be possible to work backwards from each individual element of the symbol structure to what it stands for and hence to work out how the symbol structure as a whole should be interpreted. According to the conception of computation at the heart of the language of thought hypothesis, any computation is ultimately a sequence of discrete transformations of symbol structures that have these two properties – and that have them, moreover, at every stage of the computation. Suppose, for example, that a

digital computer is computing some function, say the function of finding the square root of a given input. The computation will take a fixed number of steps. At any stage in the computation it is possible to halt the computation, to identify the symbol structures involved and then to interpret them – just as when one is carrying out a proof in the predicate calculus one can stop at any stage in the proof and identity the formula that one has arrived at.

It is difficult to understand artificial neural networks in these terms. One can certainly understand computation in an artificial neural network as a sequence of transformations. We can think about the spread of activation from one layer of units to another as a "step" in the computation that takes the network from an input (a particular pattern of activation in the units of the input layer) to an output (a particular pattern of activation in the units of the output layer). But this is a step of a very different sort from the steps that we can identify in a classical computation of the type envisaged by language of thought theorists. It is of course possible to take a "time-slice" of a network, so that one could examine what is going on in a particular layer of hidden units at a given time. However, this would not really tell one anything about what the network is doing. What the time-slice reveals is a particular pattern of activation across the units in the relevant layer, but this will not bear any sort of intuitive relation to the pattern of activation in the units of the input and output layers.

We can make this concrete by returning to the example of the mine–rock detector network (Gorman and Sejnowski 1988) considered in Chapter 5. Recall that the task of this network is to take an acoustic "fingerprint" for an underwater sonar echo and then to classify it as either coming from a rock or a mine. The fingerprint of the sonar echo is given by the distribution of energy levels across 60 different frequencies. This fingerprint is fed into the network by dedicating each unit in a 60-unit input layer to representing the energy level of the sonar echo at a particular frequency. Suppose, for the sake of argument, that we accept that the pattern of activation across the input units represents the sonar echo. There is a sense in which this can be described as a symbol structure. The pattern of activation across the layer of hidden units can for the same reasons be described as another symbol structure. We can, therefore, characterize the spread of activation from the input layer to the output layer as involving a series of computational steps – one step from the input layer to the hidden layer and one step from the hidden layer to the output layer. These computational steps are, however, very different from the steps involved in a computation defined over sentences in the language of thought. There is no obvious relation between patterns of activation at the different layers. The behavior of the hidden unit layer does not correlate in any straightforward way with the behavior of the input layer. One index of this is that very little variation is produced in the performance of the network by increasing the number of hidden units. Although the best performance came with a hidden layer of 24 units, this was a tiny

improvement over the performance in a network of 6 units.[1] The additional units are clearly not redundant, in at least the sense that the patterns of activation in the hidden layer vary significantly according to the number of units. Yet there is a huge overlap in the overall performance of these networks with fundamentally different patterns of activation in their hidden units – similar inputs lead to similar outputs, even though the intervening processes are fundamentally different. This is a long way from the classical model of computation.

Let me draw the contrast between the types of computation involved in the classical representational picture and in artificial neural networks in the starkest possible terms. The classical representational picture, most clearly developed in the form of the language of thought hypothesis, holds that cognition should be understood in terms of the rule-governed transformation of abstract symbol structures – a manipulation that is sensitive only to the formal, syntactic features of those symbol structures. That these symbol structures have the appropriate formal features is a function of the fact that they are composed of separable and recombinable components. In contrast, there are no such separable and recombinable components in artificial neural networks. The evolution of an artificial neural network takes a fundamentally different form. Since each distinct unit has a range of possible activation levels, there are as many different possible dimensions of variation for the network as a whole as there are units. Let us say that there are n such units. This means that we can think of the state of the network at any given moment as being a position in an n-dimensional space. This multi-dimensional space is standardly called the *activation-space* of the system. We can think of it as the space of all possible patterns of activation in the network. Since both inputs and outputs are themselves points in activation space, computation in an artificial neural network can be seen as a movement from one position in the network's activation space to another. There are, as we have just seen in the mine–rock network, many different trajectories between two points in the activation-space. From a mathematical point of view any such trajectory can be viewed as a vector-to-vector transformation (where the relevant vectors are those giving the coordinates of the input and output locations in activation space).

Once we start to think of the states of artificial neural networks in terms of positions in multi-dimensional activation space and the vectors that give the coordinates of those positions, it becomes clear why the notion of structure is not applicable. A point on a line does not have any structure. Nor does a point on the plane (i.e. in two-dimensional space). By extension one would not expect a point in n-dimensional space where $n > 2$ to have any structure. A similar point emerges when one thinks not of positions in the activation-space of the network but rather of the vectors that give that coor-

1 The 6-unit hidden layer had an average performance on the testing set of 83.5 percent, which was improved to 84.5 percent by quadrupling units.

dinates of those positions. A vector is simply an ordered set of numbers. It has no more and no less structure than any other ordered set of numbers. Certainly, it has nothing of the internal articulation and structural complexity of a sentence, or of a well-formed formula in a logical system.

The fundamental distinction between classical and neural networks approaches to cognition is frequently put in causal terms. The following passage from Cynthia Macdonald is a clear statement of the causal dimension of the problem:

> Both connectionist and classical models are capable of semantic interpretation. The difference is that, whereas classical models assign semantic content to expressions, i.e. symbols, connectionists assign semantic content to units, or aggregates of units, or, in Smolensky's case, to activity vectors of units. More precisely, whereas classical systems work with semantically interpretable objects (symbols) that have both causal and structural (syntactic) properties, connectionist systems work with objects (units) that have causal properties, plus objects (vectors) that have syntactic (and semantic) properties. More precisely still, the objects that causally interact at the processing level in connectionist systems are not semantically evaluable, and do not have semantically evaluable constituents. What is semantically evaluable in connectionist networks (patterns of activity or activity vectors) is *not* what does the causal work in those systems: if such systems have semantically evaluable constituents, they are, in Smolensky's words, acausally explanatory.
>
> (Macdonald 1995, p. 10)

The standard conclusion drawn from these differences is that the representational states of artificial neural networks cannot be causally efficacious. The causal work is done at the level of the individual units, while the semantic properties of particular states of the network are at best emergent properties. This standard conclusion is drawn both by those who think that the causal inefficacy of the semantic constituents of connectionist networks amounts to a reductio of connectionism (most famously Fodor and Pylyshyn 1988) and by those who think that the same phenomenon shows the inadequacy of content-based explanation (P. S. Churchland 1986; P. M. Churchland 1989b).

However, the real motivation for the standard conclusion cannot be the crude thought that, since the vectors of an artificial neural network are derived from the activation levels of individual units, the vectors themselves exist only in a derivative sense that is incompatible with their doing genuine causal work. It is hard to take seriously the thought that the existence of a causal explanation at the level of microstructure is incompatible with the existence of genuine causation at the macro-level, since this would remove at a stroke just about any type of macro-level causation. It is far more plausible to think of the standard conclusion as motivated by concerns about

structure. Let us suppose that we can map representational states onto vectors describing activation patterns in populations of units or neurons. And let us suppose, moreover, that vectors are genuinely tokened in the system. Even if there is a sense in which we can understand the output of the system as the causal outcome of vector-to-vector transformations, we still have no grip on how the causal consequences of the tokening of a given vector might be a function of the structure of its content. Vectors do not have compositional structure.

There is, then, a real tension between two approaches to the architecture of cognition. As we saw in section 9.2, proponents of the representational picture of the mind argue from very general requirements upon what it is to be a thinker to the conclusion that the vehicles of propositional attitudes must have a structure matching the structure of their contents. Yet the artificial neural networks that the neurocomputational approach takes to be our best models of the subpersonal architecture of cognition do not seem to be structured in anything like the way this argument suggests.

9.4 Rejecting the structure requirement

There is a range of ways of reacting to the tension identified in the previous section. One response is to take the tension as a compelling argument in favor of the language of thought hypothesis – as showing that artificial neural networks cannot be plausible models of the architecture of the propositional attitudes. The argument might take the following form:

1 If propositional attitudes are systematic, productive and causally efficacious in virtue of their contents, then they must have vehicles whose structure maps on to the structure of their contents.
2 If artificial neural networks are good models of the architecture of cognition, then propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.
3 Propositional attitudes are causally efficacious in virtue of their contents.
4 Propositional attitudes must have vehicles whose structure maps on to the structure of their contents.
5 Artificial neural networks are not good models of the architecture of cognition.

The argument is formally valid.[2]

2 It takes the following form:

i	$a \Rightarrow b$	
ii	$c \Rightarrow \sim b$	
iii	a	
iv	b	from (i) and (iii)
v	~c	from (ii) and (iv)

The second response accepts most of the same premises, but arrives at a fundamentally different conclusion. One might argue that, since artificial neural networks do in fact provide useful, predictive models of all sorts of cognitive abilities and since they do not appear to have the type of structure required by the language of thought hypothesis, there must be something wrong with the assumptions from which the argument begins about what it is to be a genuine thinker. The best-known version of this response is the eliminativism put forward by Paul Churchland (1979, 1981) and Patricia Churchland (1986). They take the tension between the requirement of structure and artificial neural networks as an argument against propositional attitude psychology and, more generally, against the idea that propositional attitudes have any significant role to play in cognition. Eliminativists accept the first conditional premise in the argument, namely, the premise that, *if* we really do have beliefs and desires that cause us to behave in certain ways then those beliefs and desires must be realized in physical symbol structures that have the form emphasized by language of thought theorists. They also accept the second conditional premise, maintaining that there is no room for structured sentential representations in artificial neural networks. Since their conviction that artificial neural networks provide a good model of how the brain works is stronger than their attachment to propositional attitude psychology, they derive the conclusion that there are no such things as beliefs and desires (at least as they are standardly understood). We can schematize their argument as follows:

1 If propositional attitudes are systematic, productive and causally efficacious in virtue of their contents, then they must have vehicles whose structure maps on to the structure of their contents.
2 If artificial neural networks are good models of the architecture of cognition, then propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.
3 Artificial neural networks are good models of the architecture of cognition.
4 Propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.
5 Propositional attitudes are not systematic, productive or causally efficacious.

Once again, this argument is formally valid.[3] It adds up, of course, to the conclusion that there are no such things as propositional attitudes.

3 It takes the following form:

i	$a \Rightarrow b$	
ii	$c \Rightarrow \sim b$	
iii	c	
iv	$\sim b$	from (ii) and (iii)
v	~a	from (iv) and (i)

In between these two extremes there is a range of more nuanced ways of thinking about the tension. All involve thinking critically about the two crucial premises shared between the two arguments – premises (1) and (2). There are various ways of challenging premise (1). We have already looked at some of these in previous chapters. In section 6.1, for example, we looked at Dennett's attempt to show that propositional attitudes could be causally efficacious without being realized in any sort of discrete physical structures (irrespective of any isomorphism requirement). Dennett suggests that propositional attitude explanations are made true by patterns in the behavior of intelligent agents (Dennett 1975, 1991a). These patterns are emergent properties that cannot be understood in terms of the behavior of parts of the agent. According to Dennett, the argument for structure at the level of vehicle, is based upon a confusion about the requirements of genuine explanation. In that same chapter we considered two further positions that challenge the first premise of the argument. One is the counterfactual theory of causation, according to which propositional attitude explanations are made true by the fact that certain counterfactual conditionals are true of the agent in question.[4] What makes it the case that I ϕ–ed because of my beliefs and desires is the fact that, had those beliefs and desires been different, I would not have ϕ–ed. The counterfactual theory makes the truth or falsity of propositional attitude explanations a matter solely of what goes on at the personal level. It does not impose any constraints upon the subpersonal level – and hence *a fortiori* does not impose any constraints of structure. Unlike Dennett's moderate realism and the counterfactual approach, Davidson's anomalous monism (as discussed in section 6.2) does hold that the causal efficacy of propositional attitudes stands or falls with those propositional attitudes having physically discrete subpersonal vehicles. This is the crucial step in Davidson's argument for the token-identity thesis (the thesis that each mental event is identical to some physical event). On the other hand, however, the physical events with which Davidson identifies mental events do not have the sort of structure that is demanded by premise (1) in the two arguments we have considered. Any theorist persuaded by one of these three positions will reject the first of the two premises that are common to the language of thought theorist and to the eliminativist response – and conversely, of course, nobody persuaded of the truth of premise (1) will be able to adopt any of these three positions.

These three positions are not the only way of rejecting the claim that the causal efficacy of propositional attitudes requires a structural isomorphism between vehicle and content. There are ways of developing philosophical functionalism on which the vehicles of propositional attitudes do not have sentential structure. Philosophical functionalists hold that the content of propositional attitudes is fixed by their causal role. A propositional attitude has the content it does in virtue of the causal interactions in which it typ-

4 See section 6.3.

ically enters – the states of affairs in the world that typically give rise to it; the effects it has within the propositional attitude system; and its possibilities for combination with other propositional attitudes to cause behavior. Philosophical functionalists envisage an isomorphism between a network of law-like generalizations defining causal roles at the personal level and an isomorphic pattern of subpersonal states occupying those causal roles. It is not part of the position, however, that those subpersonal states should themselves be structured in the manner suggested by the argument we are considering. This means that the problem of structure can work both in favor of and against philosophical functionalism. Theorists convinced by the claim that causal efficacy requires structured vehicles will argue that philosophical functionalism has to accept that the occupants of propositional attitude roles have to be structured, in which case it collapses into a version of the language of thought theory. But, on the other hand, it is open to those doubtful of the requirement of structure to show that philosophical functionalism can be developed in ways that secure causal efficacy without structured contents.

A version of this second strategy has been adopted by David Braddon-Mitchell and Frank Jackson, who have suggested a map-based model of cognitive architecture as an explicit alternative to the language of thought hypothesis. According to the mental maps model (Braddon-Mitchell and Jackson 1996), the vehicles of propositional attitudes are quasi-pictorial representations of the states of affairs being thought about. As in the language of thought theory, the idea of structural isomorphism is central, but it is developed in a very different way. Mental maps are supposed to be isomorphic with what they represent. The relations (or at least some of them) holding between elements of the mental map can be mapped on to the relations holding between objects in the represented state of affairs. In this way representation is secured through the relations of exemplification and resemblance. The mental map represents a state of affairs by exemplifying that state of affairs' structure – that is to say, by itself possessing a structure that resembles (at some suitable level of abstraction) the structure of the represented state of affairs. The structure of the map cannot, however, be separated out from the representational properties of what it represents in the way that the structure of a sentence can (whether that sentence is in English, the language of thought, or the first order predicate calculus). Although a map is a structured entity, its structure cannot be formally specified. Braddon-Mitchell and Jackson put the point clearly:

> There is no natural way of dividing a map at its truth-assessable representational joints. Each part of a map contributes to the representational content of the whole map, in the sense that had that part of the map been different, the representational content of the whole would have been different. Change the bit of the map of the United States between New York and Boston, and you change systematically what the map says. This is

> part of what makes it true that the map is structured. However, there is no preferred way of dividing the map into basic representational units. There are many jigsaw puzzles you might make out of the map, but no single one would have a claim to have pieces that were all and only the most basic units.
>
> (ibid., p.171)

We need, therefore, to distinguish weak and strong senses in which a representational vehicle might be structured. In the weak sense there is structure whenever a structural isomorphism can be identified between the vehicle and what it represents. In the strong sense, however, structure requires the existence of basic representational units combined according to independently identifiable combinatorial rules. Natural language sentences (or for that matter sentences in the language of thought) are clearly structured in the strong sense, whereas mental maps/models only possess structure in the weak sense.

The mental maps hypothesis has not been worked out in anything like the detail of the language of thought hypothesis, but it is relatively easy to see where the potential worries are going to arise. A defender of the language of thought hypothesis is likely to stress the intimate relation between inference and structure explored in earlier sections. There is a sense in which mental maps are structured, since they contain elements that can feature in further mental maps. Nonetheless, it is far from clear that they are structured in the right sort of way to permit the types of inference built into propositional attitude psychology. It is easy to see how there could be some very basic forms of inferential transition between maps. There might, for example, be associations between mental maps, allowing one mental map to give rise to another map, or to some particular form of behavior. The possibility of such transitions would enable maps to serve as guides to action. However, those very features of maps (their analog nature and structural isomorphism with what they represent) that make them so useful for guiding action do not allow these transitions to be viewed in inferential terms. In order to think of one map as entailing another, or as making it more probable, or as requiring a particular course of action, the maps must be interpreted in propositional terms. We have to interpret one map as expressing one proposition and the second as representing a further proposition, and then evaluate the inferential relations (be they deductive, inductive or probabilistic) between those two propositions. Once again, the language of thought theorist is likely to object, our only understanding of how to do this rests upon the two propositions being linguistically formulated.

Braddon-Mitchell and Jackson do not directly address this issue, but they do offer the following explanation of how maps can evolve over time in what is clearly intended to be an analogy with inferential transitions between linguistic representations:

> Maps are physical entities whose structure can govern the way they evolve over time. When cartographers update maps or put two maps together to make one that incorporates all the information in a single map, these operations are governed in part by the structures of the maps they are working on. And in order to find a target, rockets use a kind of internal map that gets continually updated as new information comes in. In these rockets, later maps are causal products of earlier maps plus what comes in via the rocket's sensors. Hence map theorists can tell an essentially similar story to language of thought theorists about how thoughts evolve over time as a function of their propositional objects.
>
> (ibid., p.173)

This is unlikely to convince language of thought theorists, however. For them the issue is not really about how thoughts evolve over time. In a very important sense individual thoughts quite simply do not evolve over time. It is systems of thought that evolve, and they do so as a function of the inferential relations between the thoughts that compose them. I might acquire a new belief, for example, because it is entailed by some beliefs that I already have – or, conversely, I might revise one of my beliefs when I come to appreciate its inconsistency with other things that I believe. The real problem is not understanding how I acquire or reject beliefs, but rather understanding the inferential relations between thoughts that partially explain why I acquire and reject beliefs. These inferential relations hold between distinct thoughts and nothing that Braddon-Mitchell and Jackson say in this short passage gives us any way of understanding how we should understand inferential relations between distinct thoughts at the level of mental maps. The process of combining maps has only very limited analogies with the process of inferring one thought from another. We do not, for example, have any idea what a conditional map might look like – and consequently little understanding of how conditional reasoning might take place at the level of mental maps.

Again, however, it is relatively clear how the map theorist will respond. The aim of the mental maps theory, it will be pointed out, is not to attempt to provide a subpersonal model for accounts of inferential transitions between thoughts. Mental maps are not intended as implementations of the representational view of the mind. It is true that, if we think about the vehicles of propositional attitudes as mental maps, then we cannot think of the mind as a digital computer – as performing formally specifiable operations on syntactic objects. But that is hardly a *reductio* of the mental maps theory, since Braddon-Mitchell and Jackson are trying to offer an alternative to thinking about the causal relations between mental states in terms of formally specifiable operations on syntactic objects. The real question is whether the alternative is satisfactory on its own terms – that is to say, whether it does justice to the complexity of the relations between propositional attitudes that seem to be implicated in our models of psychological

explanation. At the moment the mental maps theory has not been sufficiently worked out for it to be clear how to go about answering this question (particularly when one takes into account the problems, pointed out in Chapter 3, with the idea that the content of a propositional attitude can be fixed solely as a function of its causal role). Nonetheless, the mental maps theory provides a suggestive counterbalance to arguments that the subpersonal vehicles of propositional attitude contents must be sententially structured.

9.5 Finding structure in artificial neural networks

Section 9.3 looked at two arguments, each drawing a very different conclusion from the same two fundamental premises. One is an argument for the language of thought hypothesis, while the other is an argument for rejecting propositional attitude psychology. It is striking that two such wildly contrasting positions can be derived from a jointly held pair of premises. The first of these premises is that the causal efficacy of propositional attitudes requires a structural isomorphism between vehicle and content. The previous section explored a range of ways of challenging this premise. Let us turn to the second premise. This is the claim that artificial neural networks cannot be structured in anything like the manner required by the first premise.

A natural way of responding to this claim about artificial neural networks would be to try to identify a sense in which artificial neural networks can be structured. But there is a fundamental objection to any such attempt to identify structure in connectionist networks. Jerry Fodor and Zenon Pylyshyn argue that the search for structure in artificial neural networks is pointless (Fodor and Pylyshyn 1988). Suppose that researchers do in fact succeed in showing that artificial neural networks have the kind of compositional structure required by the first premise in the argument. Suppose, that is, that it is shown that artificial neural networks, contrary to initial appearances, do allow an isomorphism between content and vehicle. This means that we will be able to identify, within a given network, elements corresponding to the different components of a thought – elements that can combine to form further thoughts, and so on. What this will show, according to Fodor and Pylyshyn, is that artificial neural networks really just offer a way of *implementing* the language of thought – as opposed to being an alternative cognitive architecture. If the artificial neural network really can do everything required by the language of thought hypothesis, then we can abstract away from the details of the individual units and the spreading patterns of activation and consider it purely and simply in terms of sentences in the language of thought. It is clear, after all, that sentences in the language of thought will have to be implemented in something. When we talk about sentences in the language of thought, we are really describing the high-level functional organization of the brain – without any real sense of how that high-level functional organization should properly be characterized at the

neural level. If the functional organization of artificial neural networks turns out to be describable in the same terms, then all that will be revealed is an interesting model of how sentences in the language of thought might be realized in the brain. This would strengthen rather than weaken the language of thought theory. So, according to Fodor and Pylyshyn, connectionist approaches to cognitive architecture confront a fatal dilemma. Either they will fail to demonstrate the required level of structure in artificial neural networks, or they will simply reveal artificial neural networks to be implementations of the language of thought.

The only way to meet this powerful challenge is to show that artificial neural networks can be structured in a way that meets the requirements of causal efficacy without simply collapsing into implementations of the language of thought hypothesis. The most worked out and discussed proposal in this area has come from Smolensky (Smolensky 1988, 1991, 1995), who has drawn attention to a class of connectionist representations (*tensor product representations*) that are compositionally structured in a way that approximates to the structure to be found in a language of thought architecture while nonetheless remaining sufficiently different not to count simply as implementations of the sort of architecture required by the language of thought hypothesis.

The challenge for the connectionist, as Smolensky sees it, is to show how a set of structured objects (such as propositions) can be mapped onto a vector space in a way that preserves the constituency relations within each structured object and allows constituents to feature in more than one structured object. On the one hand, the mapping is supposed to preserve enough features of the propositions for it to be legitimate to describe the mapping as structure preserving. But on the other the propositional structure is not so comprehensively captured in the network that the network can be described as a mere implementation of the language of thought hypothesis. The strategy Smolensky adopts has three stages.

Smolensky begins by proposing a simpler way of representing the structure of a proposition. He deploys the technique of vector decomposition to break complex structured items down into sets of pairs by analyzing the item in terms of a set of roles each occupied by a particular filler. In the LISP programming language developed by John McCarthy, propositions are represented by binary branching trees that lend themselves particularly clearly to role-filler decomposition. To take the example developed in Smolensky (1991), the proposition *Sandy loves Kim* is represented in LISP by the following tree (Figure 9.1).

Each branch on the tree can be represented as a role with a particular filler. The leftmost branch is obviously the predicate role occupied by the – *loves* – filler. The rightmost branch corresponds to the two "gaps" in the predicate role and is filled by the ordered pair <*S*, *K*>, each of which occupies a branch of the relevant sub-tree.

The second step in the mapping is to assign primitive vectors to the roles

(33)

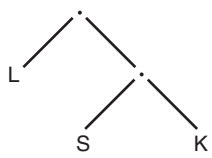

Figure 9.1 Tree for *Sandy loves Kim* (source: McDonald and McDonald (1995)).

and fillers. Let R_0 and R_1 be the role vectors corresponding to the two branches of any given node. These vectors are independent of each other (and hence cannot be multiples of each other). Let *L, S* and *K* be the filler vectors corresponding to – *loves* –, *Sandy* and *Kim*. The third step in the mapping is to define the operations that combine these filler-vectors and role-vectors into a complex vector representing the proposition *Sandy loves Kim*. There are two such operations: superposition (vector addition) and the tensor product (a complex form of vector multiplication). We can symbolize these by '+' and '@' respectively. It is the tensor product operation that binds fillers to roles. The tensor product of two vectors *V* and *W* is the vector containing all possible products of one element of *V* with one element of *W*. Thus:

$$V \otimes W = (V_1W_1, V_1W_2 \ldots V_iW_j \ldots)$$

If there are *n* elements in *V* and *m* elements in *W* then the tensor product *V* @ *W* will have *nm* elements.

The tensor product representation of *Sandy loves Kim* will take the following form:

$$R_0 \otimes L + R_1 \otimes [R_0 \otimes S + R_1 \otimes K]$$

The explanation is as follows. At the highest level of organization the proposition is a binary node with two branches, the left one occupied by the filler *L* and the rightmost one by the filler composed of the ordered pair <*S*, *K*>. The initial binding is, therefore, of the *L* vector to the vector representing the left branch role R_0 – that is, $R_0 \otimes L$. Superposed on this constituent vector (by the operation of vector addition) is the vector representing the right-hand branch, itself composed of a sub-tree with left-hand and right-hand branches occupied respectively by the vectors *S* and *K*. This particular tensor product is recursive – the object occupying the overarching R_1 role is itself represented by a tensor product vector, namely, $[R_0 \otimes S + R_1 \otimes K]$.

Tensor product vectors of this type satisfy two principal desiderata for structure-sensitive processing. First, and most obviously, they allow structured objects (such as propositions) to be mapped onto connectionist networks even though those networks do not contain discrete objects corresponding to the components of the structured objects. Hence they offer

a sense in which connectionist networks can be compositional. Tim van Gelder has reminded us of the distinction, highly relevant in this context, between two types of compositionality – *concatenative compositionality* and *functional compositionality* (van Gelder 1990). A representational system is compositional in the concatenative sense when it represents a structured item in a way that preserves tokens representing constituents of the item in the representation of the structured item and that is sensitive to how the complex item is built up from the simpler items. Concatenatively compositional systems of representation include natural languages and the various formal languages employed in mathematics, computer science, and so on. A representational system is functionally compositional, on the other hand, just if it can represent structured objects in such a way that there are effective and reliable processes for producing such an expression given its constituents and decomposing the expression back into its constituents. All concatenative compositional systems are functionally compositional, but the converse does not hold. Connectionist systems that support tensor product representations are prime examples of functionality without concatenativity. The compositional structure of complex objects can be coded into such systems via vector addition and vector multiplication, and subsequently recovered via the techniques of vector decomposition – even though the tensor product representations themselves do not preserve tokens representing the constituents of the appropriate structured objects.

It is not clear, however, that functional compositionality fully meets the content causation constraint. There is a difference between processing that preserves structure (or more accurately, processing from whose starting-point and end-product structure can be recovered) and processing that is structure-sensitive. This is where the second feature of tensor product representations comes into play. Tensor product representations go beyond mere functional compositionality to allow for a degree of recombinability. This emerges from the distinction between filler and role. A given role can be occupied by a range of different fillers and a filler that was formerly in one role can reappear in a different role. Hence there is an important sense in which tensor product representations can do more than simply represent constituent structure. Despite being vectors that encode distributed representations they can represent constituent structure in a way that allows the same constituent to feature in a range of complex objects and to occupy different roles within those complex objects.

Does the combination of these two feature give us a way of satisfying the content causation constraint without causal isomorphism and a language of thought? Fodor and Pylyshyn think not (Fodor and Pylyshyn 1988). They accept that Smolensky's tensor product framework does allow the coding of constituent structure, but argue that it cannot properly accommodate the systematic nature of thought. The systematicity of thought places strong demands upon the ability to exploit the structure of individual thoughts. These demands have been formulated in different ways. Gareth Evans

provided one well-known formulation in his 1982 book *The Varieties of Reference*. According to Evans's Generality Constraint, a subject can only properly be described as having the thought that *a is F* if he is capable of thinking the thoughts *b is F* for any object *b* of which he has an appropriate concept and *a is G* for any property *F* of which he has an appropriate concept. Although the Generality Constraint is only applicable to thoughts expressible in subject-predicate form, a more global version has been proposed by George Rey who suggests that, for any compositionally structured thought *p* that a thinker is capable of thinking, that thinker will be capable of thinking all the thoughts whose content is fixed by any logical permutation of the logico-syntactic parts of *p* (Rey 1995).

It is not under dispute that a network obeying some form of the systematicity requirement can be developed using the tensor product framework. The problem is that systematicity, in the eyes of Fodor and Pylyshyn, is not an accidental but an essential feature of any cognitive system:

> No doubt it is possible for Smolensky to wire a network so that it supports a vector that represents *aRb* if and only if it supports a vector that represents *bRa*; and perhaps it is possible for him to do that without making the imaginary units explicit (though there is so far no proposal about how to ensure this for arbitrary *a*, *R*, and *b*). The trouble is that, although the network architecture permits this, it equally permits Smolensky to wire a network so that it supports a vector that represents *aRb* if and only if it represents a vector that represents *zSq*.
>
> (Fodor and Pylyshyn 1988, in McDonald and McDonald 1995, p. 216)

The best that the tensor product framework can do, according to this objection, is to produce models of cognitive systems that are contingently systematic, whereas any cognitive system that is genuinely to qualify as a thinking system must be necessarily systematic.

At this point the argument appears to have gone full circle. Recall that we began with what appears to be a serious problem for any attempt to model the architecture of cognition using artificial neural networks. Artificial neural network models face a dilemma. If they fail to accommodate certain fundamental features of thought then they will *ipso facto* be disqualified as serious models. But if, on the other hand, they do succeed in accommodating those features of thought, then it looks as if they will not really be alternatives to the language of thought hypothesis. We can now see how this general problem works out in detail. I suggested earlier that the only way to escape the dilemma for the artificial neural networks theorist was to show how neural networks can approximate to a given characteristic of the language of thought. This is effectively to undercut the second horn of the dilemma, by showing how a neural network can reflect central characteristics of compositional thought without being a mere implementation of a language of thought architecture. Smolensky's tensor product approach

attempts to do this. However, as we have seen, it is very difficult to strike the correct balance. If the approximation is too approximate, then it is open to the language of thought theorist to object that the target has been missed completely. This is effectively Fodor and Pylyshyn's charge with respect to systematicity. The tensor product networks do not, it is true, count as mere implementations of a computational architecture – but that is only because they fail to provide a genuine sense of systematicity.

Nonetheless, we should not be too hasty to give the victory to the language of thought theorists. One obvious question to ask is whether thought really is systematic in the way that Fodor, Pylyshyn and many others have assumed without argument. There are very real questions to ask about how and why we should take it as a *datum* that thought is systematic. What sort of evidence might there be for the idea that thought is systematic? What is the status of the generality constraint, and other related requirements of systematicity? Is the idea of systematicity really as unproblematic and uncontentious as language of thought theorists tend to make out?

The obvious place to begin is with natural language. Natural languages clearly have a range of conspicuous and well-understood combinatorial features. These combinatorial features are an obvious, and acknowledged, model for those who think about the systematicity of thought. In particular, we can identify two key assumptions. The first assumption is that natural languages are systematic in the very sense in which thought is being claimed to be systematic. It is widely held that something like the Generality Constraint holds for language, in such a way that nobody could properly be described as understanding a sentence of the form '*a* is *F*' unless they were capable of combining the predicate expression '– is *F*' with other proper names in their vocabulary to form new sentences, and similarly of exploiting the proper name '*a*' in sentences that use the range of other predicate expressions in their vocabulary. One way of justifying this constraint upon linguistic understanding is that it marks the difference between the genuine understanding of a language and what is often called phrase book understanding. The basic thought is that one can only properly be described as understanding a language if one can understand how sentences are built up from their constituent words – and one can only understand how sentences are built up from their constituent words if one is able to put different combinations of words together to form new sentences.

Once the idea that natural languages are systematic is clearly in view, a second assumption comes into play. This second assumption is that the systematicity of natural language is derived from, and explained by, the systematicity of thought. Natural languages are systematic because it is their role to express thoughts. This justifies us in working backwards from the systematicity of natural languages to the systematicity of thought. There is, of course, a particular view of the relation between thought and language at stake here. This is what is sometimes called the communicative conception of language. According to the communicative conception of language, the

nature of thought can be understood independently of the nature of language. Language does not have a role to play in structuring thought. Language serves only to communicate thoughts that can be completely understood independently of their linguistic expression. Nonetheless, we can use the expressive power of language as a guide to the expressive power of thought – on the plausible assumption that the system of thought must be at least as expressively powerful as the language that is required to express it.

Once these two assumptions are in the open, it is natural to wonder whether they are compulsory. There seem to be a number of places where questions might be raised. We might start right at the beginning, with the basic model of the systematicity of natural language. It seems very clear that formal languages, such as the predicate calculus, are systematic in the very straightforward way reflected in the language of thought hypothesis and in Evans's generality constraint. The first-order predicate calculus has predicate names ('*F*', '*G*', '*H*', and so on) and object names ('*a*', '*b*', '*c*', and so on) and is governed by rules that make it the case that any formula in which an arbitrary predicate name is applied to an arbitrary object name will count as a well-formed formula. It is natural, therefore, to think that something like the generality constraint must be true of the predicate calculus, so that nobody can properly be described as understanding the predicate calculus unless they understand that any predicate name can in principle be combined with any object name – and indeed unless they understand that what it is for something to play the role in the predicate calculus of an expression that names an object is that it should be capable of being concatenated with the name of any predicate to form a sentence. But why should one think that natural languages are like this? Is it really the case that I cannot properly be described as understanding an arbitrary name, say the numeral '9' as a name of the number 9, without being able to understand any sentence that can be formed by concatenating the numeral '9' with any predicate that is in my vocabulary? There is some plausibility in the view that part of what it is for me to understand that '9' refers to 9 is that I have no idea what to make of sentences such as "9 is fat and lazy" – which is exactly the opposite of what the generality constraint appears to prescribe. After all, understanding a sentence is at least in part a matter of understanding what it would be for that sentence to be true (understanding that sentence's truth-conditions), and by the same token understanding a name is understanding how that name contributes to the truth-conditions of sentences in which it features. But, arguably, the sentence "9 is fat and lazy" does not have any truth-conditions. There is no state of affairs that is the state of affairs of the number 9 being fat and lazy – and part of what it is to understand that '9' refers to 9 is precisely to understand that there is a huge range of predicate expressions with which it does not make any sense at all to combine the numeral '9'.

One might wonder whether '9' is unique in this respect, or whether there might be a more general phenomenon here. Suppose we use the phrase *range*

of application for the predicates that it makes sense to combine with a given name, and the range of names that it makes sense to combine with a given predicate. Perhaps every name and predicate has a restricted range of application. Perhaps, moreover, we cannot separate out understanding any given predicate or name from understanding its range of application – or, at the very least, from understanding what sort of principles might circumscribe its range of application (in the way that the principle that the numeral '9' names an abstract object means that we cannot combine it with predicates applicable to concrete objects). If this is right, and it certainly has some intuitive plausibility, then nothing like the generality constraint could possibly be correct as formulated and, as a consequence, one might well wonder how secure our intuitions are about the systematicity of natural language.

A language of thought theorist is likely to think that reflections such as these completely miss the point. It would be a mistake, at least as far as the language of thought hypothesis is standardly developed, to think of the language of thought on the model of a natural language. Quite the contrary. The language of thought is generally conceived to be much more like a formal language than a natural language. As a consequence, the language of thought (or Mentalese) lacks some of the quirks of ordinary natural languages, such as the quirk of only permitting a limited degree of systematicity. One motivation for introducing the language of thought hypothesis is the idea that a language of thought is required to explain certain facts about linguistic comprehension, such as the fact that we are capable of disambiguating ambiguous sentences in particular contexts and the fact that we are capable of correctly identifying the logical form of natural language sentences. It looks as if we need a tool for thinking with that is much more precise than natural language – and, in fact, if it is indeed the case that we need a tool for thinking with that can represent the logical form of natural language sentences, then (on certain widely held assumptions about the logical form of natural language) it seems to follow that the language of thought will look very similar to the predicate calculus.

A similar conclusion follows from the metaphysical claims that are made on behalf of the language of thought theory. Recall that the language of thought hypothesis is proposed as a way of resolving the problem of causation by content – the problem, that is, of explaining how the way that a belief represents the world can be causally efficacious in generating further beliefs and/or behavior. The language of thought hypothesis is claimed to solve the problem because of the fact that the semantic properties of sentences in the language of thought are carried in the syntax of those sentences. Causal transitions between sentences in the language of thought track the semantic and logical relations holding between the contents of those sentences. But the acknowledged inspiration for this way of thinking about sentences in the language of thought in both syntactic and semantic terms is certain meta-logical characteristics of formal systems – in particular, the soundness and completeness of the first-order predicate calculus. The

further away one moves from thinking about the language of thought as a formal system, the less plausible this picture becomes.

There is, then, room for considerable debate about the relation between the systematicity of formal systems, the systematicity of thought and the systematicity (or lack of it) of natural languages. But underlying this debate is a far deeper issue, to which we have already adverted in describing the language of thought hypothesis as committed to the communicative conception of language. Part of what is at stake in the debate between the language of thought and artificial neural networks approaches to cognitive architecture are some very fundamental assumptions about the nature of thought and language, and the relation between the two. According to the language of thought hypothesis, the systematicity of natural language is a function of the systematicity of thought. Natural language is a vehicle for communicating thoughts, and it needs to be systematic because the thoughts that it has to communicate are systematic. But why, one might ask, should the order of explanation take this direction? Why should we be so confident that the structure of the language we speak has no role to play in determining the structure of the thoughts that we can think? Many theorists, both psychologists and philosophers, have found it plausible that the range of thoughts we are able to think is a function of the means we have at our command for formulating and expressing them. What is distinctive about the language of thought hypothesis is the idea that each individual needs a private language, or idiolect, in order to be able to think. The obvious question to ask, however, is why this additional step is required. Why should we assume that the language upon which the capacity to think depends is a private inner language, as opposed to a natural language?

Even granting the points made earlier to the effect that natural languages are not *perfectly* systematic in the way that formal languages can be systematic, our intuitions about the systematicity of thought still seem to derive largely from intuitions about the systematicity of language. It is difficult to think about the systematicity of thought except through the compositional structure of the sentences that express the relevant thoughts. Accordingly, one might wonder (at least as far as the requirements of systematicity of thought are concerned) whether the work that the language of thought is called upon to do could not be done by a suitably internalized natural language. (The qualification is important, since the language of thought hypothesis is also brought in to solve metaphysical problems about mental causation, and it is not so clear that a natural language could solve these problems.) It is worth exploring the possibility, therefore, that the language of thought is a natural language – that, to the extent that our thinking does need to have linguistic vehicles, the language in question is a natural language, acquired in the normal course of human development and without any peculiar formal or meta-logical properties. On this view there is nothing particularly mysterious about the linguistic dimension of cognition. As we learn a language, we acquire new modes of thought, as a function both of new vocabulary and of new methods of putting words together to form sentences.

We do not use language to express pre-existing thoughts. Rather, the language that we possess both circumscribes and defines the thoughts that we are able to think.

The proposal that the language of thought is a natural language is a compromise position. On the one hand it concedes many of the points about the need for systematicity and structure made by proponents of the language of thought hypothesis. It clearly entails, for example, that some thoughts have linguistic vehicles composed of recombinable elements. Hence it is incompatible with any views, such as some of the more extreme pronouncements of the Churchlands, holding that it is always a mistake to look for linguaform vehicles for thoughts. On the other hand, however, it can be viewed as far less of a global hypothesis about cognitive architecture than the language of thought hypothesis as standardly developed. The language of thought hypothesis comes as part and parcel of the representational approach to the mind and is closely associated with the picture of the mind as a digital computer. There is much more at stake in the representational picture than the relatively circumscribed issue of how we view the vehicles of personal-level propositional attitudes, because the representational picture is wedded to a much broader view of how information is processed in the mind/brain more generally. According to Fodor and other language of thought theorists, the language of thought does far more than simply provide subpersonal vehicles for propositional attitudes. One indication of this is that the language of thought is supposed to provide the cognitive architecture for both modular and non-modular processes.

In a sense, therefore, the proposal that the language of thought is a natural language could be seen as a drastic rescaling of the explanatory pretensions of the language of thought hypothesis as standardly conceived. The proposal is perfectly compatible with the idea that the predominant cognitive architecture in the mind is connectionist in form. It is perfectly possible to combine the idea that artificial neural networks do in fact provide accurate models of the vast majority of cognitive abilities with the further thought that some types of high-level cognition involving propositional attitudes require linguistic vehicles. As long as one thinks that these linguistic vehicles must be sentences in a private language of thought that is more akin to a formal language than a natural language, then it looks as if the second thought will be in conflict with the first – because one might reasonably expect the private language of thought to play a significant role in cognition more generally. But once one starts to think of the vehicles of propositional attitudes as being sentences in a natural language, there is far less temptation to identify a more global explanatory role for subpersonal vehicles of that type. The way is open for a more two-tiered approach to cognition and to the architecture of cognition. One might think, for example, of the higher forms of cognition associated with propositional attitude psychology as complex cultural artifacts that are superimposed upon many layers of more primitive cognitive abilities. It is the presence of natural

language that makes this superimposition possible, but this does not mean that we should think of all cognition as being essentially linguistic in form. It might well be that those cognitive abilities that do not involve propositional attitudes can be fully understood without assuming that they involve any structured, language-like representations – or, at least, without assuming that they involve representations that are more structured and language-like than one might find in networks such as Smolensky's tensor product networks. Many of these might be cognitive abilities that emerged relatively early in the course of evolution and that are shared with non-human animals – cognitive abilities that many theorists would think are particularly suited to being modeled in the terms characteristic of artificial neural networks. Moreover, it may well be that (as suggested in Chapter 7) propositional attitudes are far less widely implicated in cognition than is standardly thought. In which case it would look even less plausible to apply the linguaform model across the board in thinking about cognition.

9.6 Overview

In this chapter we have been considering how to respond to a powerful argument deployed by supporters of the language of thought hypothesis. This is the argument that genuine thought is an activity that must involve the manipulation of structured objects that can be put into a one–one correspondence with the logical structure of the sentences that express the content of the relevant thoughts. Let us call this the *structure requirement*.

The structure requirement can be embedded in two further arguments that reach conclusions diametrically opposed to each other. One argument reaches a substantive conclusion about cognitive architecture, namely, that the architecture of cognition must be that proposed by the language of thought hypothesis and the representational picture of the mind. The second argument arrives at the conclusion that propositional attitude psychology is fundamentally misconceived and that it is a mistake to think of propositional attitudes as being causally efficacious at all. This second argument is really a form of eliminativism, on the plausible assumption that saying that propositional attitudes are not causally efficacious is tantamount to saying that there are no such things as propositional attitudes.

Here are the two arguments again.

Argument 1

1 If propositional attitudes are systematic, productive and causally efficacious in virtue of their contents, then they must have vehicles whose structure maps on to the structure of their contents.
2 If artificial neural networks are good models of the architecture of cognition, then propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.

3 Propositional attitudes are causally efficacious in virtue of their contents.
4 Propositional attitudes must have vehicles whose structure maps on to the structure of their contents.
5 Artificial neural networks are not good models of the architecture of cognition.

Argument 2

1 If propositional attitudes are systematic, productive and causally efficacious in virtue of their contents, then they must have vehicles whose structure maps on to the structure of their contents.
2 If artificial neural networks are good models of the architecture of cognition, then propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.
3 Artificial neural networks are good models of the architecture of cognition.
4 Propositional attitudes cannot have vehicles whose structure maps on to the structure of their contents.
5 Propositional attitudes are not systematic, productive or causally efficacious.

The two arguments share premises (1) and (2). They both accept the structure requirement, and they both accept that artificial neural networks do not satisfy the structure requirement. Where they differ is on the weight they respectively attach to the idea that artificial neural networks provide good models of the architecture of cognition.

There are ways of thinking about the causal and explanatory role of propositional attitudes that are clearly incompatible with the structure requirement. Some of these are closely linked to different ways of developing what we have called the picture of the autonomous mind. According to theorists such as Dennett, for example, the causal efficacy of propositional attitudes does not depend upon their having discrete inner vehicles. Propositional attitudes are emergent properties of the cognitive system as a whole. A similar view is taken by theorists who adopt a counterfactual approach to mental causation and hold some version of the view that a particular complex of propositional attitudes causes an action just if, had the agent not had those attitudes, she would not have performed the action in question. Some versions of the autonomy picture do allow for (and indeed require) causally efficacious inner items. Davidson's anomalous monism is a case in point. But anomalous monism does not require those inner items to be structured in the way proposed by the structure requirement. The structure of the physical events that are identical to beliefs and desires is a function of how those events are described, not of their intrinsic nature. There are ways of developing the functional picture of the mind that are equally incompatible

with the structure requirement. In section 9.4 we looked at the theory of mental maps proposed by Braddon-Mitchell and Jackson. The mental maps approach allows for a degree of structure in the vehicles of propositional attitudes, but one that falls far short of that demanded by the structure requirement.

It is, furthermore possible to challenge the second premise, even in the face of Fodor and Pylyshyn's powerful argument that any artificial neural network that satisfies the structure requirement will *ipso facto* count as simply an implementation of the language of thought. In section 9.5 we looked at Smolensky's tensor product networks, which arguably provide an approximation to the structure requirement that nonetheless falls short of being a mere implementation of the language of thought.

The final suggestion that emerged in section 9.5 was that the structure requirement might not be a global requirement upon the architecture of cognition in the way suggested by proponents of the language of thought hypothesis and of the representational picture of the mind. It might be confined to a relatively small part of cognition – to what in Chapter 8 was described as the core of the cognitive system, namely, those cognitive abilities and capacities that involve the propositional attitudes. If this is the case, then the language of thought hypothesis and artificial neural networks need not necessarily be viewed as offering competing accounts of the architecture of cognition, but rather as applicable to different aspects of cognition. Moreover, if the range of application of the structure requirement is circumscribed in this manner, then the possibility opens up that it might be satisfied without assuming a private, internal and proto-formal language of thought. Perhaps the requisite structure could somehow be derived from the structure of natural language. This possibility raises some very fundamental questions about the relation between thought and language. These questions will be pursued in the next chapter.

10 Thinking and language

- Thinking in words (1): the inner speech hypothesis
- Thinking in words (2): the rewiring hypothesis
- The state of play
- Practical reasoning and the language of thought
- Perceptual integration
- Concept learning

Key points in debates about cognitive architecture are closely bound up with issues about the linguistic nature of thought. According to the language of thought hypothesis, cognition must be linguaform, as a consequence of what are taken to be certain very basic facts about the nature of thought and representation. The "master argument" for the language of thought hypothesis is the argument that the *systematicity* and *generativity* of thought can only be explained if thinking essentially involves the manipulation of sentence-like structures. The previous chapter explored certain aspects of this argument. Although the chapter focused primarily on the debate between language of thought and connectionist approaches to cognitive architecture, we briefly considered the possibility of a compromise position that would derive the systematicity and generativity of thought from the systematicity and generativity of a natural language. This chapter pursues this idea further, exploring the dialectic between language of thought theorists and those who think that natural languages can do all the work that the language of thought has been called upon to do.

There is an important set of issues about the "direction of fit" between public language and thought. So, for example, a basic plank of the case for the language of thought hypothesis is that one needs a language to learn a language. We cannot, it is claimed, acquire a public language unless we already have at our disposal a language for formulating hypotheses about what words mean. This argument incorporates both a specific empirical claim about the mechanics of language learning and a philosophical claim about what it is to understand a language. As we shall see, both claims can be disputed. Related to this is a more general claim about how language is used to communicate. The language of thought hypothesis sits very naturally with what is sometimes called the *communicative* conception of language, public language is simply a tool for the communication of ideas. The fact that we are participants in a public language does not have any implications for the structure and content of our thoughts. Rather, it is the structure and content of our thoughts that give meaning to the sentences that we

use, because the intentions that we have in using language are what determine the way it is understood. Opponents of the communicative conception, on the other hand, hold that there are fundamental differences between the cognitive capacities of language-using creatures and the cognitive capacities of non-linguistic creatures. Participation in a public language makes available types of thinking that would otherwise be inaccessible. Language has a structuring role to play in cognition.

In section 10.1 we will consider what I call the *inner speech hypothesis*, according to which we think in the words of a natural language. Section 10.2 explores an extension of the inner speech hypothesis. This is the rewiring hypothesis to the effect that the acquisition of a public language (both in the development of the species and the development of the individual) effects a fundamental change in cognitive architecture, making available types of thinking that are simply not available in the absence of language. As emerges in section 10.3, the obvious response that supporters of the language of thought hypothesis might make to the inner speech and rewiring hypotheses is that they leave us without the resources to explain thinking behavior in non-linguistic creatures. In addition to the "master argument" from generativity and compositionality discussed in the previous chapter Fodor has a powerful line of argument to the effect that certain very basic cognitive abilities are language-dependent. This is an argument for the language of thought hypothesis because these basic cognitive abilities can plausibly be ascribed to many creatures that lack a public language. The remaining sections of the chapter explore Fodor's arguments for the language-dependence of these basic abilities. Section 10.4 considers practical decision-making, which Fodor takes to involve computations of expected utility and hence a linguistic medium in which those computations can be performed. Sections 10.5 and 10.6 explore the domains of perceptual processing and concept learning. Fodor thinks that these both take the same form, involving the formation, testing and refining of hypotheses. These hypotheses, whether they are hypotheses about objects in the distal environment or about the extensions of concepts, need to be linguistically formulated. Again, since it is implausible to think that perception and concept learning are confined to language-using creatures, Fodor concludes that there must be a language of thought. Section 10.6 also considers Fodor's further argument that the very process of language learning requires a representational medium with at least the expressive capacity of the language being learnt.

10.1 Thinking in words (1): the inner speech hypothesis

According to the language of thought hypothesis, all thinking involves manipulating sentence-like structures that display an isomorphism between their syntactic and their semantic properties, so that the structure of the content of the thought is reflected in the structure of the physical object that actually enters into causal transactions within the cognitive system. This

picture of how the mind works is claimed to have two fundamental advantages. First, it is supposed to explain *causation by content* in virtue of the isomorphism between syntax and semantics (see sections 4.1 and 4.2 above). Second, it is claimed to be the only way of explaining our ability to think indefinitely many new thoughts and to understand permutations of thoughts that we are currently entertaining.

Proponents of the language of thought hypothesis have a further proposal about the nature of those sentence-like structures. They hold, as we explored in the previous chapter, that these sentence-like structures must be sentences in an internal language of thought that is independent of any public language. It is this further proposal that makes the language of thought hypothesis so distinctive and that crystallizes many of the central claims that language of thought theorists make about the relation between thought and language. We can explore these claims by considering how a language of thought theorist might respond to an obvious challenge. Even if one grants the need for thinking to have sentence-like vehicles, why cannot these sentence-like vehicles simply be sentences of a public language? Why do we need to think in a private internal language of thought? Why cannot we think in and through a public language?

The force of this challenge depends upon how the proposed alternative is understood. What does it mean to say that we think in and through a public language? The most radical proposal in this area is the *inner speech hypothesis* (Sellars 1969; Carruthers 1996). This is the idea that our conscious thinking (the type of thinking that we engage in when we respond to questions, set out to solve problems and deliberate about what to do) involves explicitly manipulating the sentences of a public language. According to the inner speech hypothesis, we can think of propositional attitudes as relations to public language sentences that are silently uttered or entertained in thought. On this view, propositional attitudes end up looking rather similar to speech acts, with belief being construed for example as a type of internalized assertion and the process of deliberation coming out as a type of inner monologue.

The inner speech hypothesis applies only to *propositional thinking* – to the types of thinking that we describe using the vocabulary of the propositional attitudes. We engage in various types of *non-propositional* thinking. There are certain types of problem that we solve by manipulating mental images and exercising the visual imagination. We are conscious of our own bodily sensations, emotional feelings and other such qualitative states. Moreover, as stressed in previous chapters, much of our thinking involves detecting patterns and recognizing templates. None of these are examples of propositional thinking. When we use our visual imagination to calculate whether the parking space is wide enough for the car, or whether the backhand shot will remain in play, we are not contemplating propositions but rather manipulating visual images.

There are two parts to the distinction between propositional and non-propositional thinking. The first has to do with the content of the thoughts.

Propositional thoughts are thoughts with contents that can be reported and expressed in 'that –' clauses. For each propositional thought there is a sentence that we would intuitively accept as giving its content. Things are very different when it comes to the various types of non-propositional thinking. Here it is much harder, and in fact usually impossible, to find a sentence that gives the content of, for example, our visual imaginings. We can give a general indication of what we are thinking about by saying that we are calculating whether the car will fit into the parking space – just as we can give an indication of what we are perceiving by describing what is in front of us. But we cannot find, in the case of visual imagination any more than in ordinary perception, a single sentence that will come anywhere near to capturing the way we are thinking about the world. Almost all the details of the scene and of our individual perspective on it will inevitably be left out. What is characteristic of propositional thinking, in contrast, is that there is nothing more to the content of a belief, say, than what is captured in the sentence that gives its content.

This is connected to the second difference between propositional and non-propositional thinking. Propositional thoughts can be evaluated for truth or falsity and the truth-value of one thought can be related to the truth-value of another thought in a way that allows us to make inferences from one thought to another. These inferences can be deductive or probabilistic – that is, they can tell us what must be the case if a particular thought is true, or what is likely to be the case if that same thought is true. The ideal rational thinker is one who makes transitions between thoughts that mirror the logical relations holding between those thoughts. Nothing like this holds in the case of non-propositional thinking, however. When I try to work out whether my car will fit into the parking space, I may entertain a sequence of images of the car being parked. But there are no logical relations holding between these images. It is not the case, for example, that an image of my car alongside the parking space entails or even makes probable an image of my car safely parked in the space. Nor is it appropriate to speak of a logical or probabilistic relation holding between a complex social situation and the pattern that is extracted from that situation.

With the distinction between propositional and non-propositional thinking in mind, we can ask whether we are ever introspectively aware of propositional thoughts that are *not* in the form of public language sentences. It is not hard to find examples of mental events and conscious states that are not inner public language sentences – the difficulty comes in making the case that any of these count as propositional. Are we ever acquainted with propositional thoughts that are not already "clothed" in the words of a public language? Defenders of the inner speech hypothesis think not, typically appealing to introspective evidence.

There is a natural objection to any such appeal to introspection. After all, we have formulated the distinction between propositional and non-propositional thoughts in such a way that the key characteristic of a

propositional thought is that it should be capable of being put into words. So, if we are to identify a thought as propositional, we will need to put it into words. But then, in the very act of identifying a propositional thought *as a propositional thought* we will have created an inner sentence, and so it is not surprising that we have the impression that there could not be propositional thoughts that are not clothed in the form of a public language sentence.

But this objection may concede too much. The defender of the inner speech hypothesis claims that we are only aware of propositional thoughts that come in the form of a public language sentence. This claim is not being significantly challenged. The opponent of the inner speech hypothesis needs to establish that we can be aware of thoughts that are both non-linguistic and propositional. It seems plausible, however, that one cannot be aware of a thought that is propositional without being aware of its content, namely, without being aware of the proposition that it expresses. But then the question immediately arises of what the vehicle of that content could be. What is it that we apprehend in a wordless form and then put into words?

The contortions that one gets into when one tries to answer this question are evocatively brought out in some important passages from Wittgenstein's *Philosophical Investigations* (1953). Wittgenstein is exploring the very natural idea that we can get at what a thought is through the differences between what goes on when we utter a sentence out loud with understanding and what goes on when we utter words that we do not understand – we can think of the thought as what it is that we grasp and try to convey when we utter a sentence with understanding, on the assumption that the meaning of a sentence is the content of the thought that it expresses. "Is thinking a kind of speaking?" he asks, continuing:

> One would like to say it is what distinguishes speech with thought from talking about thinking. – And so it seems to be an accompaniment of speech. A process which may accompany something else, or can go on by itself.
>
> (ibid., §330)

The picture is a natural one. Surely something must be going on when we understand language – some kind of mental action that is independent of the words we actually utter, and that gives them meaning. Once we grant that much, it seems only a short step to the idea that that mental action, whatever it is, could take place without there being any words at all (either publicly uttered or silently uttered). Suppose we describe that mental action as thinking the thought that is expressed by the sentence we utter (or that we might utter). Since this thought would typically be a propositional thought, it seems to follow that we can think without thinking in words – that we can apprehend a wordless thought.

Wittgenstein suggests that this apparently inescapable conclusion is deeply problematic. The difficulties emerge in the following passage:

> While we sometimes call it "thinking" to accompany a sentence by a mental process, that accompaniment is not what we mean by a "thought". — Say a sentence and think it; say it with understanding. — And now do not say it, and just do what you accompanied it with when you said it with understanding!
>
> (ibid., §332)

It is hard to see what one could do except either to think the original sentence to oneself again or to think of another sentence that says the same thing in different words. The temptation (if there is one) to think that it *must* be possible to carry out Wittgenstein's instruction is most likely to come from the thought that there are all sorts of occasions when we seem to find ourselves fitting words to thoughts in ways that suggest that we have an independent grasp of the thought and can check how accurately the words match up to it. Wittgenstein devotes considerable effort to trying to show that these frequent and familiar occasions are better described in rather different terms.

> What happens when we make an effort – say in writing a letter – to find the right expression for our thoughts? — This phrase compares the process to one of translating or describing: the thoughts are already there (perhaps were there in advance) and we merely look for their expression. This picture is more or less appropriate in different cases. — But can't all sorts of things happen here? — I surrender to a mood and the expression *comes*. Or a picture occurs to me and I try to describe it. Or an English expression occurs to me and I try to hit upon the corresponding German one. Or I make a gesture, and ask myself: What words correspond to this gesture? And so on.
>
> (ibid., §335)

Wittgenstein is offering us different ways of describing what is going on in cases where we might find it intuitive to appeal to wordless thought. Lying behind the specific redescriptions is a diagnosis of what has gone wrong. The problem, he thinks, is an illicit move from the obvious and correct thought that there must be something that makes it the case that we use language with understanding to the far more problematic thought that there must be something *accessible to conscious introspection* that makes it the case that we use language with understanding. His alternative proposal is that using language with understanding is a matter of participating in a public practice. What makes it the case that a sentence is uttered with understanding is not something going on in the mind of the speaker at the time of uttering the sentence, but rather to be found in what leads up to the sentence and what happens after it – the situation (which might be linguistic or non-linguistic) to which the sentence is a response and how the speaker is disposed to continue to act (once again, where the action might be either linguistic or non-linguistic).

It is not clear that Wittgenstein himself is a proponent of the inner speech hypothesis, although he comes close to it in passages such as the following: "When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought" (ibid., §329). But it is not difficult to see how Wittgenstein's ideas could be deployed to support the inner speech hypothesis. The more we chip away at the idea that there might be wordless thoughts, the more plausible it becomes to hold that the vehicles of conscious propositional thinking are natural language sentences.

Yet the inner speech hypothesis has its own difficulties. Consider the following version of the inner speech hypothesis put forward by Peter Carruthers: "We mostly think (when our thinking is conscious) by imaging sentences of natural language, and trains of thought consist of manipulations and sequences of such images" (1996, p. 228). Carruthers talks about "imaging" a natural language sentence. What does this mean? Is entertaining a public language sentence in thought comparable to hearing a public language sentence? Is it an acoustic matter? It is hard to see what it might mean to think of one's hearing being directed inwards, or how there could be a sound that makes no noise. And, even if we could make any sense of this type of internal audition, it is still puzzling how one could actively manipulate a sound token in the manner required if one is actively to think, rather than have thoughts occur to one. Nor is it any easier to think of entertaining a public language sentence as analogous to seeing an inscription of a public language sentence (although it is easier to understand what it might be to manipulate such an inscription). But if our access to an internalized public language sentence is neither auditory nor visual, then it is hard to see what explanatory power we have gained by talking about inner sentences at all. It looks as if we will have to talk, not about the inner entertaining and manipulating of a public language sentence, but rather about the inner entertaining and manipulating of some sort of *representation* of a public language sentence. And it might well be thought that this leaves us right back where we began, with having to explain the nature of this representation. If what Carruthers describes as an imaged public language sentence is really the representation of a public language sentence, then the question of what the vehicle of that representation is remains completely unexplained.

This brings us to a second difficulty with the inner speech hypothesis. The inner speech hypothesis is a hypothesis about the vehicles of conscious propositional thinking – of the type of reasoning that typically involves a succession of occurrent judgments that come in a more or less logical order. It has little to say about the architecture of cognition more generally. In contrast, the language of thought hypothesis is put forward as a general model of the mechanics of all types of thinking – and, it might be suggested, among the types of thinking that the language of thought can explain are precisely those types of thinking that are highlighted by the inner speech hypothesis. Even if we grant the principal arguments in

support of the inner speech hypothesis (namely, the direct argument from introspection and the indirect arguments put forward by Wittgenstein and others), it could still be argued that these tell us only about the *phenomenology* of thinking. In other words, the inner speech hypothesis can only give us a first-person perspective on the nature of thinking; on how it seems to the subject rather than on how it is from a third-person point of view.

There is a fundamental difference between an account of what thinking is like from the point of view of the thinker and an account of what makes it possible for there to be thinking at all. We have already seen reasons for thinking that the inner speech hypothesis does not offer us an account of the second type. It does not explain what it is that allows us to "image" natural language sentences, given that we cannot understand this "imaging" in any sort of straightforward perceptual manner (as a type of inner hearing or inner vision). It may be that the first-person experience of silently thinking about public language sentences is only possible because the relevant public language sentence is represented in the language of thought. The proposal might be that what is really going on when we engage in inner speech is that we represent to ourselves a public language sentence, where the vehicle of that representation is a sentence in the language of thought.

It is important to keep these questions about phenomenology and cognitive architecture apart. No argument from phenomenology is likely to be persuasive against the language of thought hypothesis, since the language of thought hypothesis is not making a claim about phenomenology. Recall the distinction between personal and subpersonal levels of explanation that we began with in Chapter 2. The phenomenology of thinking is a personal-level phenomenon, whereas the language of thought hypothesis is a subpersonal-level account of how the mind works. The language of thought theorist can accept all the substantive claims of the inner speech hypothesis. The language of thought theorist does not have to claim (and would be advised not to claim) that conscious propositional thinking involves manipulating sentences in the language of thought in any sense that would imply that we are introspectively acquainted with sentences in the language of thought. Such a theorist would be much better off holding that it is the manipulation of sentences in the language of thought at a subpersonal level that grounds the introspectively accessible personal-level phenomenology of inner speech.

The structure of the debate here mirrors a well-known debate in cognitive science about the nature of visual imagery. A number of experiments have offered powerful evidence that certain types of problem solving seem to involve manipulating visual images – problems that involve rotating shapes or imagining how things look from a completely different perspective (Shepard and Cooper 1982; Wraga and Kosslyn 2003). So, for example, in one well-known set of experiments, Roger Shepard presented subjects with pairs of three-dimensional shapes and asked them to determine whether one shape was a rotation of the other. It turned out (Shepard 1982) that the time taken to answer the question varied in direct proportion to the extent to

which one shape was an angular displacement of the other – a result that might naturally be taken to suggest that in some sense subjects are solving the problem by rotating the shapes and seeing whether they map onto each other. These experiments have given rise to a lively debate about whether cognitive information processing involves depictive representations (where a depictive representation is one that bears a pictorial resemblance to what it represents, in the way that a painting or a map does). Supporters of the language of thought hypothesis have tried to show that the experimental data can be accommodated without postulating depictive representations at the subpersonal level. Supporters of imagistic representations have disagreed. But the issue here is not about the types of thinking that are introspectively accessible at the personal level. Both sides in the debate can agree that we sometimes have the experience of imaginatively rotating shapes and transforming images. The real issue is about the type of representations that have to be postulated at the subpersonal level in order to explain how and why we do have that experience.

So, it is open to the language of thought theorist to try to take on board the points made by the inner speech hypothesis and show how her account of cognitive architecture can explain and accommodate them. In this sense, then, the inner speech hypothesis is not in direct conflict with the language of thought hypothesis. The two accounts are pitched at different levels of explanation. However, there is a natural extension of the inner speech hypothesis that *is* incompatible with the language of thought hypothesis. This is what I will call the *rewiring hypothesis*.

10.2 Thinking in words (2): the rewiring hypothesis

Suppose, for the sake of argument, that conscious, propositional thinking does involve the manipulation of public language sentences. Does this have any implications at the level of cognitive architecture? The language of thought and the inner speech hypotheses are not directly in competition, but there is a way of developing the basic idea at the heart of the inner speech hypothesis (the idea that we think in and through a public language) that does directly challenge the language of thought hypothesis. The central claim of the *rewiring hypothesis* is that there are fundamental differences between the cognitive architectures of language-using and non-language-using creatures. The development of language in human pre-history served to rewire the human brain in ways that create fundamental differences between the types of thinking available to linguistic and non-linguistic creatures. This process of rewiring is recapitulated in individual development as the human infant acquires language. According to the rewiring hypothesis, the acquisition of language (in both *phylogeny* and *ontogeny*) reconfigures the cognitive architecture of the brain, making available new types of representation and computation.

The core of the rewiring hypothesis is a conception of the mind/brain as a complicated structure of mechanisms and circuits superimposed on top of

each other. The mind/brain is the product of many thousands of years of evolution, and evolution is a process of modification and tinkering. Some mechanisms disappear in the course of evolution. Others are modified beyond all recognition. Still others persist even though some of their functions have been taken over by newer and more specialized mechanisms. It is natural to think of the mind/brain as containing a huge range of specialized circuits and mechanisms, of different levels of sophistication and with different evolutionary lineages.

Neuroanatomists frequently make a standard and very broad distinction between three phylogenetically distinct compartments of the human brain. The most primitive part is protoreptilian, and composed of the spinal cord and areas such as the basal ganglia that are thought to be involved in procedural learning and motor skills. The next compartment in terms of evolutionary history and sophistication is the so-called *limbic system,* generally thought of as paleomammalian (dating back to the earliest history of mammals) and implicated in memory, emotions and the motivation of behavior. The most evolutionary recent parts of the brain are in the neocortex, generally thought to be the home of various higher cognitive functions. Each of these cerebral compartments is responsible for different aspects of human behavior and everyday life involves a constant switching from one compartment to another (and, of course, comparable switching within compartments). As new regions, areas and circuits evolved, they had to accommodate themselves to what was already there. And this of course was a two-way process. Older areas and circuits were modified and transformed by the newer areas and circuits grafted on to them.

We need to view the rewiring hypothesis against the background of this picture of the brain as a complex structure of mutually adapting and interconnected mechanisms and circuits. The theme that emerges when one thinks about the evolution of the brain is one of flexibility and plasticity. Circuits that originally evolved for one function are recruited to new functions. New connections are made between different areas. What type of stimuli does it take to initiate and continue this complex process of evolution and adaptation? No doubt many of these changes are due to significant events in human pre-history, such as the descent from the trees or the gradual shift to living in larger and larger social groupings (Donald 1991; Mithen 1996). Many such changes fit the standard pattern of evolutionary explanation. A change in circumstances poses a problem, to which certain mechanisms are better suited than their competitors. The result is selection for the genes that code for those mechanisms – and so the mechanisms gradually take their place in the genotype of the species and in the phenotypes of individual members of that species.[1]

When one thinks about the evolution of the brain it is most natural to

1 The genotype of a species is the set of genetic instructions coded in its DNA, while the phenotype is the individual organism that results from the interaction between genotype and environment.

think about it in the terms just outlined. This is how evolutionary psychologists tend to think about the evolution of the brain. Consider the rationale for the cheater detection module postulated by Cosmides and Tooby.[2] The cheater detection module is supposed to have evolved in response to a specific problem confronted by our Pleistocene ancestors, namely, the need to be able to identify those who have taken a social good without providing the corresponding social benefit. The background assumption is that social interactions in human pre-history were regulated by something like the TIT-FOR-TAT heuristic, which tells one (roughly speaking) to cooperate with anyone who has not reneged in the recent past. The evolutionary rationale for the cheater detection module is to identify the renegers, and the cheater detection module is claimed to have evolved because it increased the fitness of those individuals carrying the "cheater detection" genes. The end result was that the genes for the cheater detection module became incorporated in the genotype.

Whatever one thinks about explanations and hypotheses of this type, it seems clear that they cannot be the only explanation of phenotypical changes. The point is well made by Daniel Dennett in his book *Darwin's Dangerous Idea*:

> We often make the mistake of confusing a cultural innovation with a genetic innovation. For instance, everybody knows that the average height of human beings has skyrocketed in the last few centuries. (When we visit such relics of recent history as *Old Ironsides*, the early-nineteenth-century warship in Boston Harbor, we find the space below decks to be comically cramped – were our ancestors really a race of midgets?) How much of this rapid change in height is due to genetic changes in our species? Not much, if any at all. There has been time for only about ten generations of *Homo sapiens* since *Old Ironsides* was launched in 1797, and even if there were a strong selection pressure favoring the tall – and is there evidence for that? – this would not have had time to produce such a big effect. What have changed dramatically are human health, diet, and living conditions; these are what have produced the dramatic change in phenotype, which is 100 percent due to cultural innovations, passed on through cultural transmission: schooling, the spread of new farming practices, public-health measures, and so forth. Anyone who worries about "genetic determinism" should be reminded that virtually all the differences discernible between the people of, say, Plato's day and the people living today – their physical talents, proclivities, attitudes, prospects – must be due to cultural changes, since fewer than two hundred generations separate us from Plato.
>
> (1995, p. 338)

2 See section 8.4.

Why should the point that Dennett makes about height and other physical features not apply even more clearly to the human brain? Perhaps the structure of the brain is just as influenced by cultural factors as are simpler physical phenomena such as height and life expectancy. Might there well be cultural factors that have played a role in the forming the architecture of cognition?

When one thinks about the cultural changes that are most likely to have had a significant effect on the development of the architecture of cognition, the most obvious candidate is the emergence of a public language. The central claim of the rewiring hypothesis is that the emergence of a public language, even though it took place in relatively recent evolutionary history, has resulted in a fundamental change in the way that the brain processes information – a fundamental change, not just in how we think and communicate about the world at the personal level, but also in how the brain processes information at the subpersonal level. The rewiring hypothesis does not simply think of language as a tool that allows us better to organize and communicate our thoughts, as well as to take short-cuts in picking up skills by being able to profit from the experience and advice of others.[3] The rewiring hypothesis is the equivalent at the subpersonal level of the inner speech hypothesis at the personal level. Just as the inner speech hypothesis holds that we can only engage in conscious, propositional thinking in and through the sentences of a public language, the rewiring hypothesis holds that certain types of information processing are only possible as a function of the cerebral rewiring that comes with the emergence of language. There is a "step-change" between the linguistic brain and the non-linguistic brain – although, of course, the linguistic brain is superimposed upon the non-linguistic brain and does not completely take over and co-opt the functions and mechanisms of the non-linguistic brain.

Unsurprisingly, the evidence for the rewiring hypothesis is largely indirect and, skeptics would say, highly circumstantial. Some suggestive material comes from the study of early hominids by cognitive archeologists (Donald 1991; Mithen 1996; Mellars 1996).[4] The consensus among archeologists and students of human evolution is almost universal that the crucial stage in human cognitive evolution occurred about 40,000 to 35,000 years ago, with the transition from what is known as the Middle Paleolithic to the Upper Paleolithic. This transition involved a sudden explosion in tool technology and social/cultural organization, with the emergence for the first time of forms of life that are recognizably congruent with those of modern humans. It is here that we find the first decorative objects; the first really compelling evidence for totemistic/religious behavior, as revealed in burial practices and totemic representations; sophisticated hunting strategies that capitalize on seasonal

3 See Andy Clark (1997, Chapter 10, and 1998) for very suggestive discussion of the prosthetic functions of language.

4 For overviews, see Donald (1991, Chapter 8), Mellars (1996) and Mithen (1996, Chapter 9).

migrations and fluctuations in animal numbers; and far more complex forms of tool production that seem to have drawn upon detailed knowledge of natural history to tailor tools for particular hunting tasks. From this point on the rate of cognitive evolution accelerated exponentially. It is tempting (and many cognitive archeologists have succumbed to the temptation) to see this transition as involving the emergence of a recognizably human language.

Even supporters of the rewiring hypothesis recognize, however, that suggestive correlations such as these are of no use without some concrete proposal as to how the emergence of a public language can make available new types of thinking and new types of information-processing. Although theories in this area are inevitably going to be highly speculative, it is worth drawing attention to two interesting lines of thought that have been put forward. One line of thought is pitched at the subpersonal level, and suggests that the distinctive contribution of natural language to cognitive architecture lies in providing a representational medium for integrating different forms and types of information (Carruthers 2002). A second has been developed at both subpersonal and the personal levels. This is the idea that public language provides a medium whereby a cognitive system/thinker (depending on whether one is considering matters at the subpersonal or personal levels) can explicitly represent its own representations/thoughts in a way that makes them available for further processing/thinking (Karmiloff-Smith 1992). Let us look briefly at each of these in turn.

Many psychologists and cognitive scientists have suggested that much cognition is domain-specific and modular. Theorists have proposed bodies of knowledge and correlative mechanisms specialized for processing information about, for example, numbers; the dynamic and kinematic behavior of objects; the mental states of other subjects; the properties and characteristics of living objects; and, of course, the detection of free-riders. These domain-specific modules are held to exist in adult humans, human infants and non-human animals – that is to say, in both language-using creatures and non-language-using creatures. Some have suggested, however, that the crucial difference between language-using and non-language-using creatures is that only the former are capable of integrating the information from different domain-specific modules (Mithen 1996, Chapter 10; Carruthers 2002; Bermúdez 2003a, Chapter 9). Language offers a medium for recoding domain-specific representations in a way that will allow them to be integrated with each other. This suggestion does allow us to make sense of one interesting feature of the archeological record. Early hominids appear to have been unable to integrate their practical abilities in tool construction with their detailed knowledge of natural history. Archeologists have inferred this from their failure to produce handaxes for specific purposes. In the Middle Paleolithic, for example, we find what seem to be highly developed tool-making skills existing side by side with a subtle and advanced knowledge of the natural environment, but it is not until the Upper Paleolithic that we see these two bodies of knowledge being integrated in the form of

tools such as fish-hooks and bone harpoons, together with hunting strategies that are tailored to the habits of specific animals (Mithen 1996).

But what is it about language that allows it to serve as a means for integrating different types of domain-specific information? Here is one hypothesis. It seems plausible that different domains of "knowledge" are represented in different ways, reflecting the different ways in which they are acquired and the different functions that they serve. One might expect the practical skills implicated in tool manufacture to be represented in a *procedural* manner, as stored motor routines that involve certain sequences of movements with clearly defined aims and expected outcomes. On the other hand, one might expect the recognitional skills required to be able to exploit the properties and characteristics of living objects to be encoded in a fundamentally perceptual format. This type of perceptual knowledge is likely to be highly modality-specific. It might involve, for example, recognitional templates for the sounds of particular animals and for the appearances of particular plants. But these highly modality-specific representations would seem to be incommensurable with procedurally encoded practical skills. A recognitional template can be used to classify perceived objects, but it is not clear how it can be manipulated and transformed in the way that it would have to be if it was to be exploited in the process of tool manufacture.

So how could an early hominid put these two bodies of knowledge together in order to adapt the design of tools to reflect specific knowledge of the natural world? To put it crudely, there needs to be some kind of common representational format that can "read" each individual representational format and allow them to communicate with each other. The hypothesis is that language can serve this function because it is a highly abstract representational medium. The knowledge built into recognitional templates is intrinsically tied to how it was acquired. It is knowledge of sights and sounds, smells and shapes. The knowledge built into the motor routines of tool manufacture is no less tied to its mode of acquisition and exercise. It is knowledge of bodily movements and how objects will respond to those movements – a combination of motor memory and perceptual expectations. But public language is a completely contrasting form of representation, since it is conventionalized and symbolic. As conventionalized and symbolic, language is amodal and does not have any immediate implications for action. It operates off-line. We can think of public language as a way of re-encoding information. The information that it re-encodes is already there in the system, but the linguistic re-encoding allows it to be used in ways that it could not otherwise be used (Karmiloff-Smith 1992).

So, the first role of language in rewiring the brain is to provide a representational medium for integrating different forms and types of information. This is closely related to the second rewiring function. Suppose we take seriously the idea that public language can be deployed to re-encode information that is already available in the cognitive system in a different format. In addition to allowing the integration of information previously in incom-

mensurable formats, this might be expected to make it possible for a cognitive system to take that information as the direct object of further thoughts. The acquisition of language does not simply provide a unified representational format that will allow different bodies of knowledge and different skills to communicate with each other; it also makes possible a new type of thinking that is explicitly directed at those bodies of knowledge and skill.

Andy Clark has made some very suggestive remarks in this area:

> Perhaps it is public language that is responsible for a complex of rather distinctive features of human thought – viz. the ability to display *second-order cognitive dynamics*. By second-order cognitive dynamics I mean a cluster of powerful capacities involving self-evaluation, self-criticism, and finely honed remedial responses. Examples would include recognizing a flaw in our own plan or argument and dedicating further cognitive efforts to fixing it, reflecting on the unreliability of our own initial judgments in certain types of situation and proceeding with special caution as a result, coming to see why we reached a particular conclusion by appreciating the logical transitions in our own thought and thinking about the conditions under which we think best and trying to bring them about … In all these cases, we are effectively thinking about our own cognitive profiles or about specific thoughts. This "thinking about thinking" is a good candidate for a distinctively human capacity – one not evidently shared by the non-language-using creatures that share our planet.
>
> (1997, pp. 208–209)

There is a powerful reason for thinking that our thoughts can only become the objects for the types of thinking characteristic of what Andy Clark calls second-order cognitive dynamics if they are linguistically encoded (Bermúdez 2003a). Clark is talking about methods of cognitive self-monitoring – tracking inferential connections and relations of evidential support. In order to engage in second-order cognitive dynamics we need to be able to think about the logical and probabilistic connections between thoughts. We need to be able to work out when, for example, two beliefs are inconsistent with each other, or when a particular course of action seems likely to thwart our desires. But we have no understanding of logical and probabilistic relations between thoughts except in so far as those thoughts are linguistically formulated. Logic and the probability calculus (and hence, by extension, decision theory and other formal theories of rational choice) track relations between sentences.

There are two different (but not exclusive) ways of developing this version of the rewiring hypothesis. It can be developed at either the personal or the subpersonal levels. At the personal level, the claim is that we can only engage in conscious and reflective cognitive self-monitoring through the medium of public language. This is, in effect, a narrow version of the inner speech hypothesis (from which it differs in not saying anything about

ordinary, first-order thinking). More ambitiously, these ideas might be developed at the subpersonal level. The idea here would be that cognitive systems that do not participate in public language could not engage in self-monitoring of any type. They are not capable of monitoring connections between thoughts even at the subpersonal level. We can appreciate the implications of this version of the rewiring hypothesis by considering what it rules out. It is completely incompatible, for example, with certain widely held theories of concept learning. As we will see in section 10.6 below, many cognitive scientists have proposed that concept learning is essentially a process of hypothesis-formation and testing. On this view, we learn concepts by forming hypotheses about the types of things that fall under them. We refine these hypotheses in the light of various types of feedback, both positive and negative, until we eventually home in on a hypothesis that picks out the correct extension for the concept within an acceptable margin of error. This is not supposed to be something we engage in consciously. It is hypothesized, rather, as a series of unconscious and subpersonal processes that underwrite our personal-level abilities to apply concepts. Moreover, these unconscious and subpersonal processes are supposed to be shared with many types of non-linguistic creature that may not properly be describable as engaging in conscious reflection at all. According to the view we are considering, however, such unconscious hypothesis-formation and testing is only possible within cognitive architectures that have been "programmed" for it by participation in a public language.

It is clear that the nature of concept learning will be one of the key points at issue between proponents of the rewiring hypothesis and supporters of the language of thought hypothesis. We will be discussing it later on in the chapter. Let me end this exposition of the rewiring hypothesis, however, with another passage in which Andy Clark makes some further suggestive comments about why the availability of second-order thinking should be a function of participation in a *public* language. His comments are, I think, applicable at both the personal and subpersonal levels.

> In order to function as an efficient instrument of communication, public language will have to be molded into a code well suited to the kinds of interpersonal exchange in which ideas are presented, inspected, and subsequently criticized. And this, in turn, involves the development of a type of code that minimizes contextuality (most words retain essentially the same meanings in the different sentences in which they occur), is effectively modality-neutral (an idea may be prompted by visual, auditory, or tactile input and yet be preserved using the same verbal formula), and allows easy rote memorization of simple strings. By "freezing" our own thoughts in the memorable, context-resistant, modality-transcending format of a sentence, we thus create a special kind of mental object – an object that is amenable to scrutiny from multiple cognitive angles, is not doomed to alter or change every time we are exposed to new inputs or

> information, and fixes the ideas at a high-level of abstraction from the idiosyncratic details of their proximal origins in sensory input. Such a mental object is, I suggest, ideally suited to figure in the evaluative, critical, and tightly focused operations distinctive of second-order cognition.
> (Clark 1997, p. 210)

The suggestions Andy Clark makes in this passage are explicitly directed at public languages. Linguistically formulated thoughts are the only possible vehicles of second-order cognitive dynamics because critical reflection can only be directed at thoughts that have certain properties – properties such as abstractness, fixity of meaning, amodality and being relatively insensitive to context. These properties are in turn derived from the communicative role of language. Language needs to be abstract, amodal, context-insensitive and to have relatively fixed meanings if it is to serve as a tool for communication.

10.3 The state of play

Sections 10.1 and 10.2 explored two different (but not exclusive) ways of developing the general idea that the language of thought is a natural language. According to the inner speech hypothesis, the vehicles of conscious thought are the sentences of a natural language, and propositional thinking is a matter of manipulating those public language sentences. The inner speech hypothesis is not in itself incompatible with the language of thought hypothesis. A language of thought theorist can grant the inner speech hypothesis as a hypothesis about the phenomenology of thinking, while holding that we still need a language of thought to explain the subpersonal information-processing that makes it possible for us to have the personal-level experience of consciously manipulating public language sentences. The real tension with the language of thought hypothesis comes with the rewiring hypothesis, which holds that there are certain aspects of cognition and cognitive architecture that can only be understood in terms of a public language. The development of language in human pre-history is claimed to have rewired the human brain in ways that create fundamental differences between the types of thinking available to linguistic and non-linguistic creatures. This process of rewiring is recapitulated in individual development as the human infant acquires language. The end result (in both phylogeny and ontogeny) is a reconfiguration of the cognitive architecture of the brain, making available new types of representation and computation.

The rewiring hypothesis offers an alternative to the argument for the language of thought from the systematicity and the generativity of thought. There has to be a language of thought, the argument runs, because without a language of thought we would be unable to explain our abilities to grasp an indefinite number of new thoughts by recombining the individual elements of thoughts that we are already able to think. The rewiring hypothesis,

however, suggests a way of showing how the systematicity and generativity of thought can emerge from participation in a public language. Among the many changes in cognitive architecture that occur when one learns a public language (or perhaps more accurately, when one *grows into* a public language) is the emergence of a representational tool that makes it possible, through the manipulation of sentences and their parts (at the personal or at the sub-personal level), to formulate and explore the implications of an indefinite range of thoughts. So too with the other features of thought for which the language of thought is claimed to be a necessary condition. We should view them not as fixed features of thinking *per se*, but rather as properties of a distinctive type of thinking that itself only emerges as a product of language.

Once the issue is put in these terms, it is clear how the language of thought theorist will reply. The hypothesis that the language of thought is a public language is a hostage to empirical fortune. It stands or falls with the idea that there is a significant fault line running through the animal kingdom marking the distinction between the sophisticated types of cognition available only to language-users and the more primitive types of cognition that are the lot of all other species besides *Homo sapiens* (as well, of course, as those members of *Homo sapiens* who have yet to acquire a language). And this is an idea with which the language of thought theorist is likely to have very little sympathy. Consider the following passage from Fodor:

> The obvious (and, I would have thought, sufficient) refutation of the claim that natural languages are the medium of thought is that there are nonverbal creatures that think. I don't propose to quibble about what is to count as thinking, so I shall make the point in terms of the examples discussed in Chapter 1. All of the three processes that we examined there – considered action, concept learning and perceptual integration – are familiar achievements of infrahuman organisms and preverbal children. The least that can be said, therefore, is what we've been saying all along: Computational models of such processes are the only ones we've got. Computational models presuppose representational systems. But the representational systems of preverbal and infrahuman organisms surely cannot be natural languages. So either we abandon such preverbal and infrahuman psychology as we have so far pieced together, or we admit that some thinking, at least, isn't done in English.
>
> (1975, p. 56)

Unlike the master argument for the language of thought hypothesis, there is no mention here of systematicity and generativity. The argument is much more basic. Fodor claims that there are certain very basic forms of cognition that would quite simply not be available in the absence of some sort of linguaform representational system. These basic forms of cognition are widespread among creatures that do not participate in a public language, be they non-human

animals or infra-linguistic humans. It follows, therefore, that the language of thought cannot be a public language. These basic forms of cognition are:

- practical reasoning (what Fodor calls considered action in the passage quoted)
- concept learning
- perceptual integration.

The remainder of the chapter examines Fodor's reasons for thinking that each of these forms of cognition requires a linguaform representational medium. Practical reasoning will be the topic of section 10.4; perceptual integration will be discussed in section 10.5; and section 10.6 will be devoted to concept learning. Section 10.6 also considers the language of thought theorist's claim that the very possibility of learning a public language presupposes the existence of a language-like internal representational medium.

10.4 Practical reasoning and the language of thought

One motivation for the language of thought hypothesis is the claim that decision-making, and the deliberation that leads up to it, is a computational process. The details have been worked out in different ways, but the most plausible approach is decision-theoretic. On this view, which we find in Fodor's original presentation of the language of thought hypothesis (Fodor 1975) as well as in more recent defenses of the notion (Maloney 1989; Rey 1997), decision-making is essentially a matter of maximizing expected utility and an internal language of thought is required as a medium in which the relevant calculations can take place. The intuitive idea is that thinking behavior results from deliberation on the environment as represented in the light of background representations and motivational states. The decision-theoretic model is a powerful way of fleshing out this intuitive idea and, once one accepts that, it is plausible that the various components of the decision-making process (specifications of possible outcomes, calculations of preferences, assessments of probability, and so forth) will need to be represented in a language-like medium.

Fodor offers the following schematic model of practical decision-making (Fodor 1975, pp. 28–29). See also Rey (1997).

A A given creature finds itself in a certain situation S.
B It believes that a certain set of behavioral options, $B_1 \ldots B_n$, is available in S.
C The creature predicts the probable consequences of performing each of those behavioral options by computing a series of conditionals of the form: If B_i is performed in S then consequences C_i will occur with a certain probability.

D A preference ordering is assigned to the consequences.
E The creature's choice of behavior is determined as a function of the assigned preferences and probabilities.[5]

Fodor's model of practical reasoning is a descriptive psychological theory. He does not see it as providing a normative theory of rationality, since it has nothing to say about the rationality or otherwise of the creature's beliefs about possible options and outcomes; its assessments of the likelihoods of those outcomes; or its assignment of subjective utilities. But should we follow him in this?

Let us grant Fodor that calculations of expected utility do indeed require a linguaform inner representational medium. The real issue is whether we can make sense of thinking behavior (of what he calls considered action) without assuming that it involves some sort of calculation of expected utility. It may be the case that when we are thinking about an action and trying to decide whether or not it is rational we may need to look at it through the eyes of something like expected utility theory. We do often characterize people's behavior in ways that amount to saying that they are failing to maximize expected utility. We might, from a *normative* point of view, describe them as irrational on those grounds. This is something that might be said, for example, about people who take part in national lotteries. But there are two different things that one might be saying in such a situation. One might be making an *internal* judgment to the effect that the person in question made a faulty calculation. So, for example, one might be saying that they failed to notice that, however sizeable the prize in the lottery, the probability of winning is so negligible that the expected utility is effectively zero. This sort of criticism implies that they did perform some sort of expected utility calculation, but just did not do it very well. On the other hand, however, one might make an external judgment to the effect that, irrespective of how they actually decided to buy a lottery ticket, from the viewpoint of expected utility theory the decision was a bad one. This would be a judgment of the action, rather than of the details of the reasoning that led up to it.

Often when animal behaviorists and cognitive ethologists thinking about the behavior of non-linguistic creatures in terms of expected utility theory they do so from this sort of external perspective. We see a good example of

5 Fodor's model cannot be quite right as it stands. It is not (usually) the case that each behavioural option will have only one outcome, and what the creature will have to compute, for each behavioral option, are the likelihoods of the principal different outcomes that could occur. Intuitively, a creature will need to consider not simply the most desirable outcome that might be consequent upon acting in a given way, but also the less desirable and indeed positively undesirable outcomes that might also occur. Each of these outcomes will have a different utility. The decision-maker will then need to weight the likelihoods of each different outcome by its desirability. The sum of these calculations will yield an expected utility for that behavioral option. The final stage will be simply to select the behavioral option with the greatest expected utility.

this in what is known as *optimal foraging theory*, which is based on the guiding assumption that animals both should and do maximize the net amount of energy obtained in a given period of time. The calculations involved in working out the course of action that would maximize the net gain of energy are cost–benefit calculations that closely match the cost–benefit calculations of orthodox expected utility theory with acquired energy as the benefit. For a foraging bird, for example, faced with the "decision" of whether to keep on foraging in the location it is in or to move to another location, the costs are the depletions of energy incurred through flight from one location to another and during foraging activity in a particular location. The cost–benefit analysis can be carried out once certain basic parameters are set, such as the rate of gaining energy in one location, the energy cost of flying from one location to another and the expected energy gain in the new location. Optimality modeling makes robust predictions of foraging behavior in birds. Cowie's study of great tits foraging in an experimental environment containing sawdust-filled cups with mealworms hidden inside is a good example. Cowie showed that the amount of time a given bird spent at a given cup could be accurately predicted as a function of the travel time between patches and the quantity of mealworms in the cup (Cowie 1977).

Nonetheless, there is no suggestion in optimal foraging theory that birds or any other foraging animals really are carrying out complex calculations about how net energy gain can be maximized within a particular set of parameters and background constraints. It is a crucial tenet of optimal foraging theory that the optimizing behavior is achieved by the animal following a set of relatively simple rules of thumb or heuristics, which are most probably innate rather than learned. So, for example, a great tit might be hard-wired to move on to the next tree after a certain number of seconds of unsuccessful foraging in one tree. Evolution has worked in such a way (at least according to the proponents of optimal foraging theory) that foraging species have evolved sets of heuristic strategies that result in optimal adaptation to their ecological niches. This optimal adaptation can be mathematically modeled in terms of a sophisticated version of expected utility theory, but the behaviors in which it manifests itself do not result from the application of such a theory – any more than a bird's capacity to fly reflects any mastery on its part of the basic principles of aerodynamics. Here, then, we have a clear example of how something like the language of expected utility theory can be used from an external perspective, even though it is clear that no calculations of expected utility are really going on in any psychological sense. Does this not blunt the argument from practical decision-making to the language of thought hypothesis?

A supporter of the language of thought hypothesis is likely to object that the example of optimal foraging theory is not really relevant to her argument. The behavior of foraging birds is not an example of the type of thinking behavior that she is trying to explain. There is no sense in which the starling or chickadee is really *choosing* between different courses of action. The bird's behavior could simply be read off the relevant heuristics and

strategies, if we knew what they were. A successful reply to the language of thought theorist will need to show that there are ways of genuinely acting intelligently that do not involve calculating the utilities and probabilities of different possible results to arrive at a calculation of expected utility – and, moreover, will need to show that this offers us a way of thinking about the behavior of non-linguistic creatures.

Intelligent action is inextricably linked to the possibility of psychological explanation. A creature is acting intelligently when its behavior is not explicable in terms of non-psychological mechanisms, such as stimulus-response conditioning, innate releasing mechanisms or reflex responses. It is when none of these modes of explanation can be brought to bear that we find ourselves compelled to characterize a creature's behavior in terms of the way it represents its environment. Let us look more carefully at the form of such an explanation. When we are dealing with language-using creatures, the aim of a psychological explanation is to present a combination of beliefs and desires that will make it intelligible why the action in question should have occurred. The action becomes intelligible when any rational agent with those beliefs and desires and in comparable background conditions could be expected to act in the same way.

This conception of psychological explanation goes hand in hand with a particular conception of how intentional actions are generated. The basic motor of an intentional action is a desire (whether a desire for something or a desire that something be the case). But a desire alone is insufficient to bring about an action. Desires feed into action when conjoined with instrumental beliefs pointing to how those desires might be satisfied. These instrumental beliefs themselves need to be "anchored" in beliefs about the environment (as well as depending upon further background beliefs). The decision theoretic model of practical reasoning offered by Fodor and other language of thought theorists provides one way of fleshing out this standard template of belief–desire explanation. On Fodor's model the motivating desire in any particular action comes from the agent's preference-ordering, while the agent has both beliefs about the environment (in the form of beliefs about the different possible courses of action available) and instrumental beliefs (in the form of beliefs about the likelihoods of the different possible outcomes). So, all three components are clearly in place. But the interesting question is whether we can *only* do justice to the three components of psychological explanation and intentional action by using a decision theoretic model. It is far from obvious that this is the case.

There are many cases of intentional action where the relevant instrumental information is clearly contained in one's current perception of the environment, so that there is no need for an instrumental belief. If, for example, my desire is for a drink of water and I see a glass full of a liquid that looks like water in front of me well within arm's reach, then my reaching out towards the glass will not always be dependent upon a separate instrumental belief to the effect that I will be able to obtain the glass if I

reach out for it. Often I will just be able to see that the glass is within reach and act accordingly. The thesis, associated with J. J. Gibson and the ecological approach to visual perception, that the content of visual perception includes what Gibson termed *affordances* offers a way of developing this basic idea (Gibson 1979; Bermúdez 1998). An affordance is a resource or support that the environment offers a particular creature – such as the possibility of providing shelter, or the availability of food. Although affordances are relativized to particular species, so that the same region of the environment might offer different affordances to different species, they are nonetheless objective features of the environment and exist as a function of the physical properties of the environment. The basic idea behind Gibson's theory of affordances is that the environment is not perceived in neutral terms. What are perceived are the possibilities that the environment affords for action and reaction, including the potential of various locations for providing shelter, concealment or nourishment. These affordances are directly perceived in the patterns of light in the optic flow – although, of course, creatures need to become "attuned" to the relevant features of the environment. Acting upon perceived affordances is one way of acting intentionally without engaging in the complex calculations envisaged on Fodor's model. A creature might simply act upon a perceived affordance, or alternatively it might act upon one of a range of perceived affordances – the affordance of Flight, for example, as opposed to the affordance of Fight.

Defenders of the language of thought hypothesis are likely to respond that the appeal to affordances is question begging. They might concede that we can make sense of the idea of the direct perception of affordances as a feature of first-person phenomenology, but deny that this is any sort of alternative to the language of thought hypothesis. Once again the distinction between personal and subpersonal levels of explanation becomes important. It seems plausible to say that we, as adult language-using humans, are capable of directly perceiving the possibilities that the environment holds. We can just see whether an object is within reach, or too heavy to lift. It seems highly plausible also that non-linguistic creatures are capable of something similar. But this is not a brute fact about cognition. It is a personal-level feature of our conscious experience and, just like any other personal-level feature of conscious experience, it requires a subpersonal explanation. Language of thought theorists have little time for Gibson's ideas about the direct pick-up of information and about the organism "resonating" to the environment (Fodor and Pylyshyn 1981). What makes the perception of affordances seem direct to us (to the extent that we agree with Gibson about the phenomenology of perception) is that we are not aware of making any inferences or engaging in any thought about what we perceive. But this does not mean that there are no such inferences taking place at the subpersonal level. Quite the contrary. We can only perceive affordances in virtue of complex information processing. This information processing requires a representational medium, which is the language of thought.

The issues that are raised here go far beyond the question of practical reasoning with which we began. What is at stake is how we understand the nature of perception – and in particular whether the type of information processing implicated in perception requires a linguaform representational medium. This will be discussed in the next section when we turn to Fodor's argument that an internal language of thought is required for what he calls perceptual integration. As we will see there, there are alternatives to the language of thought approach to perception. In the present context this means that we need to take seriously the possibility that the subpersonal information processing involved in the direct perception of affordances may well be non-sentential form in form.

Bearing this unresolved issue in mind let me end this section by offering a further alternative to thinking about the intentional behavior of non-linguistic creatures in decision-theoretic terms. The language of thought theorist is offering a highly propositional account of decision-making. But we saw in section 10.1 that there are ways of thinking about the process of non-linguistic decision-making in non-propositional terms. We can draw a general distinction between *thinking-how and thinking-that* (by analogy with the philosophical distinction initially proposed by Gilbert Ryle between propositional knowing-how and non-propositional knowing-that). Thinking-that is a type of thinking best understood in propositional terms – where the thoughts have determinate contents that can be specified in sentences with 'that –' clauses (sentences of the form "he thinks that *p*" where *p* is a further sentence that spells out how the world has to be for the thought in question to be true). The decision-theoretic calculations that Fodor takes as paradigmatic of practical reasoning are clearly examples of thinking-that.

Here, in contrast, are four examples of *thinking-how*:

1 *Imagistic reasoning*, such as calculating whether my car will fit into a particular parking space.
2 *Trial-and-error reasoning*. Trial-and-error reasoning is driven by a representation of the goal, but often does not involve explicit hypotheses about how the goal in question is to be achieved.
3 *Analogical reasoning*. To find an analogy between two situations or two ideas is to identify a relation between them that can rarely be put into words – a relation that can be perceived, but not conceptualized in any accurate way.
4 *The exercise of complex bodily skills*. The acquisition and exercise of complex skills is a highly cognitive activity, requiring precise calibration of different types of information. But competent practitioners cannot usually express the practical knowledge that it involves linguistically.

Opponents of the language of thought hypothesis are likely to describe these as fundamentally non-propositional types of thinking. It may well be that when I try to work out whether my car will fit into the parking space I am

doing something much more like manipulating visual images than formulating conditionals about the likely consequences of different scenarios. Similarly, trial-and-error reasoning might best be described in terms of visualizing various different scenarios off-line.

The argument for the language of thought hypothesis from the requirements of practical reasoning would be blunted if it turned out that the practical reasoning required at the non-linguistic level can all be viewed as forms of thinking-how rather than thinking-that. A view of this type has in fact been put forward by Michael Dummett, who makes a distinction between what he sees as the genuine and full-fledged types of thinking only available to language-using creatures, and the *proto-thoughts* of non-linguistic creatures (Dummett 1993). One of the distinguishing features of proto-thoughts (as Dummett sees them) is that they are essentially tied to the possibilities the environment affords for action. Because of this, as the following passage makes clear, they should be seen as imaginative transformation of the perceived environment:

> The sublinguistic level of proto-thought is essentially spatial, and therefore must be conceived as operating in our apprehension of what we perceive as having a three-dimensional shape and occupying a three-dimensional position. But it is also essentially dynamic: it involves the apprehension of the possibilities and probabilities of movement, and of the effect of impact. For this reason, it incorporates, not merely perception of position, shape and movement, but also recognition of the gross properties of material things. It is an immediate feature of even our visual perceptions that we observe objects as differentiated according to the general type of material of which they consist: whether they are rigid or flexible, elastic, brittle or plastic, cohesive like a lump of sugar or a heap of grains like caster sugar, solid, liquid or gaseous, wet or dry, smooth or rough, greasy or clean, and so forth. The reason that we use visual clues to project these properties, even though unaided vision does not disclose them, is precisely that they bear on the dynamic possibilities.
>
> (Dummett 1993, p. 124)

The vehicles of proto-thoughts are much closer to perceptual states than they are to linguistically expressible propositions. According to Dummett, the vehicles of proto-thoughts are "spatial images superimposed on spatial perceptions" (ibid., p. 123). In perceiving the ambient environment proto-thinkers visualize the possible ways in which it might be transformed, drawing upon motor memories and a sense of their own possibilities for action and reaction. Proto-thinkers do not come to a judgment about what the environment contains or the possibilities it affords, where coming to a judgment implies something that can be detached from the here-and-now, but nonetheless they perceive the environment in a way that involves exercising judgment.

The language of thought theorist is not without resources at this point. She may well return to the ambiguity identified earlier between accounts of thought at the personal and subpersonal levels. It may well be that non-linguistic decision-making always involves forms of thinking-how, and it may well be that the vehicles of all thinking-how, whether engaged in by language-using creatures or by non-linguistic creatures, are imagistic. But this still does not tell us about the subpersonal cognitive architecture that underwrites thinking-how. Language of thought theorists hold that perception, and hence *a fortiori* the exercise of visual imagination and the manipulation of spatial images, requires a fundamentally propositional cognitive architecture at the subpersonal level. Since imagistic representation presupposes the language of thought it can hardly be an alternative to it. We will consider their arguments for this claim in the next section.

10.5 Perceptual integration

It emerged in the previous section that much of the plausibility of the language of thought hypothesis rests upon how the mechanisms of perception are understood. Many objections to the language of thought hypothesis try to drive a wedge between propositional thinking and non-propositional thinking, where non-propositional thinking is understood as fundamentally perceptual in form. The language of thought hypothesis attempts to block these objections by arguing that perception requires a linguaform cognitive architecture.

The basic argument here is that perception of the form that we and other sentient creatures have would not be possible in the absence of a linguaform cognitive architecture. The information provided by the sensory systems, it is claimed, grossly underspecifies how the environment appears to us in conscious perception. As a consequence, our cognitive systems are deeply implicated in constructing our perceptual representation of the world. This process of construction is best seen as a process of hypothesis-formation and testing. The final stage of the argument should by now be familiar. The computational processes of hypothesis formation and testing require a language-like representational medium, and hence a language of thought.

Fodor begins with a very abstract characterization of a sensory mechanism. A sensory mechanism, he thinks, is a device that operates "to associate token physical excitations (as input) with token physical descriptions (as output): i.e. a sensory mechanism is a device which says 'yes' when excited by stimuli exhibiting certain specified values of physical parameters and 'no' otherwise" (Fodor 1975, p. 46). The sensory mechanisms of hearing, for example, are sensitive to properties of sound waves, while the sensory mechanisms of sight are sensitive to properties of light waves. The starting-point for the argument is that the basic function of a sensory mechanism is to produce physical descriptions of the relevant properties to which they are sensitive. These physical descriptions will not themselves carry any informa-

tion about the perceived environment. The information that they carry is *proximal* (information about events that take place on the sensory periphery) rather than *distal* (information about the properties of objects independent of the perceiver). This proximal information does not feature, however, in our conscious perceptual experience of the world. We perceive the distal environment, not activity in the retina or the inner ear. So, the fundamental problem the perceptual systems have to solve is how to get from proximal information to distal information – how to derive a representation of the distal environment from information about proximal stimuli.

According to Fodor, perceptual systems solve this problem by engaging in a series of redescriptions of the initial proximal information. A given redescription at level $n + 1$ is essentially a hypothesis derived from the description at level n together with the background information available to the system. Fodor draws the following conclusions:

> If one accepts, even in rough outline, the kind of approach to perception just surveyed, then one is committed to the view that perceptual processes involve computing a series of redescriptions of impinging environmental stimuli. But this is to acknowledge that perception presupposes a representational system; indeed a representational system rich enough to distinguish between the members of sets of properties all of which are exhibited by the same event.
>
> (1975, p. 51)

Fodor is envisaging something along the following lines. Consider a situation in which two very different environmental events generate the same proximal stimuli. The perceptual system has in some sense to disambiguate the proximal stimuli to "decide" which of the two events to include in its representation of the distal environment. To this end it generates a hypothesis on the basis of the initial proximal information and whatever background knowledge is available (or, more plausibly, a series of hypotheses). This hypothesis is the result of a process of *nondemonstrative inference* (i.e. an inference that is not deductive). This inference takes as premise a description of proximal stimuli and produces as output a description of a distal event. The perceptual system can only make such an inference if it is able to represent both the physical properties that are proximally represented (patterns of sound or light waves, for example) and the physical properties that are distally represented (properties of mind-independent objects) and, of course, able to compute the probable relations between them.

In thinking about this argument it is useful to separate out two different claims. The first is that successful perception depends upon the brain being a hypothesis-testing machine. The second is that the brain can only be a hypothesis-testing machine if the process of hypothesis formation and testing takes place in the language of thought. Fodor's argument requires both claims. It is not enough simply to show that perception involves some

form of hypothesis testing. It must be the sort of hypothesis testing that requires a language of thought. Bearing this in mind we can examine the two claims separately.

The line of reasoning behind the first claim is that proximal information is insufficient to determine the final representation that emerges from perceptual processing, and therefore that significant inferential transitions are required to generate the latter from the former. This argument has a key assumption, which is that the only information that the brain has to work on in constructing a perceptual representation of the distal environment is the proximal information that arrives at the sensory periphery – the light waves impinging on the retina, for example, or the sound waves arriving at the ear. It is this assumption that generates the seemingly enormous disparity between the "raw materials" of perceptual processing and the complex three-dimensional representations that eventually emerge – and hence that makes the need for hypothesis formation and testing seem so pressing. But many psychologists would argue that proximal sensory information is just one of the inputs into perceptual processing, and hence that there is not as much disparity as there might seem between input and output in perception.

Researchers in perception stress that a considerable amount of background information about the nature of the physical world is hard-wired into perceptual systems, giving those systems a bias towards interpreting proximal information in certain ways that match up to certain fundamental characteristics of the physical world (Shepard 2001). It is clear, for example, that the perceptual systems are very sensitive to what are termed shape, size and color constancy. Our movement through the world frequently produces sudden and drastic changes in the information that the sensory systems receive from a particular object. In particular, our sensory systems have to confront sudden and drastic changes in the types of information that typically specify color, shape and size. Think, for example, of how reflectance information changes as one moves out of the shade into the sunlight, or of how patterns of shape-specifying information on the retina change as one moves around a complex object. Yet our perceptual systems smooth out these changes. As we move towards an object it occupies a progressively larger portion of the retinal image. Yet we effortlessly see an object of fixed size coming closer, rather than an object in a fixed position getting larger. How is this achieved? Many perceptual psychologists agree that assumptions of color, size, and shape constancy are built into the visual system, imposing constraints that massively restrict the number of degrees of freedom that the perceptual systems have in interpreting proximal information. It appears, moreover, that the visual system has built into it the default assumption that objects are illuminated from above – understandably, given that the visual system evolved when the sun was the principal source of illumination (Ramachandran 1988). If this is correct, then the computational challenge confronted by the perceptual systems is not quite as formidable as Fodor's argument assumes.

The existence of these hard-wired processing constraints is in one sense compatible with the argument for the language of thought hypothesis. It could be the case that interpreting proximal stimulation according to these constraints requires a language of thought. On this view, all that we have done is change the parameters of the problems confronted by the perceptual systems, rather than learning anything new about their intrinsic nature. On the other hand, however, one might think that problems of this sort are precisely the sort of problems that might best be modeled in terms of neural networks rather than computational processing – and hence that do *not* require a language of thought. There are two distinguishable issues here. The first has to do with the general form of the tasks involved – and in particular with the idea that perceptual processing is a problem that involves satisfying multiple constraints simultaneously. The second has to do with the particular form that one might expect solutions to the problems of perceptual processing to take. It is possible to argue that the type of tasks in perceptual processing involve tasks of pattern recognition and template matching best modeled by artificial neural networks.

Let us start with the first issue. In very general terms, the perceptual systems need to interpret proximal stimulation in the light of a range of constraints, including those we have identified. Neural networks are well suited to problems of constraint satisfaction (Horgan and Tienson 1996). Constraints are encoded into neural networks through the weights that attach to connections between units in different layers (recall that these weights may be excitatory or inhibitory). The weights determine the degree of influence that one unit has on another. A given pattern of weights (which might be built into the network, or "learnt" through a learning algorithm such as the back-propagation algorithm) can embed a number of different constraints. Depending on the precise form of the input into the network, one constraint might dominate a second constraint in one context and not in another. In neural networks, therefore, we have what are frequently termed *soft constraints*. Soft constraints are defeasible, rather than invariant. They operate "for the most part", rather than in the exceptionless manner characteristic of rule-governed systems. The constraints that we have identified seem to be soft constraints in this sense. Perceptual constancy is not a universal rule. Objects do change suddenly in size, shape and color. A cognitive system that is to remain suitably sensitive to such changes (and of course its survival may depend upon such sensitivity) needs to be able to override the constraints built into it.

It seems, then, that even if perceptual processing is a process of hypothesis formation and testing, it is a process that requires the type of multiple constraint satisfaction that is generally accepted to be one of the principal strengths of neural network models. This already goes some way towards weakening the argument from hypothesis formation and testing to the language of thought. But there are further and more specific reasons for thinking that neural network models are particularly appropriate for modeling

the processing that takes the perceptual systems from proximal inputs to a representation of the distal environment. These reasons arise from different ways of thinking about how best to characterize the basic tasks performed by the perceptual systems. All parties might agree that the basic function of the visual system is to generate a three-dimensional representation of the distal environment on the basis of the two-dimensional information conveyed in the retinal image. But this neutral description can be fleshed out in a number of ways. One might, for example, describe what the visual system does in computational and bottom-up terms, as process of gradually building up from the initial input in a step-by-step and rule-governed way. Describing the task in this way leads naturally to the language of thought hypothesis – to the idea that what we are dealing with is a sequence of transformations performed on syntactically specified objects. And it is of course in these terms that Fodor and other proponents of the language of thought hypothesis understand the notion of hypothesis formation and testing.

But there are other ways of thinking about how hypothesis formation and testing might take place. Let us go back to our neutral description of what the visual system does. We have agreed that its basic function is to generate a three-dimensional representation of the distal environment on the basis of the two-dimensional information conveyed in the retinal image. We might, however, see this basic task as being composed of a number of more specific and circumscribed tasks, each of which can be seen as much closer to a task of pattern recognition and template matching than to a sequence of transformations of syntactically specified objects. Let me give an example.

A fundamental problem in processing the proximal information encoded in the retinal image is generated by *binocular disparity*. Each retina receives its own pattern of stimulation and, because the eyes are some distance apart, these patterns of stimulation are significantly different. The so-called *correspondence problem* is the problem of explaining how the visual system derives a single three-dimensional image from both the two-dimensional images that differ from each other. Consider a single point in the distal environment. Light reflected from that point will typically fall on to points on both the right and left retina. If we think of each retina as a grid with locations given by fixed coordinates then we can see that light from the single point will generally fall at different locations on the two retinas. Nonetheless, these different locations on the two retinas correspond to each other, in the sense that they both receive light from a single distal source. The correspondence problem is the problem of identifying the corresponding pairs of points from the two retinas (Churchland and Sejnowski 1992, Chapter 4).

In thinking about how the correspondence problem might be solved, one plausible initial thought is that it is far easier to solve it for pairs of points that correspond to the boundaries of objects and surfaces than for pairs of points that are in the middle of homogenous surfaces. Imagine looking at a gray square against a white background. There are points in each retinal image receiving light from an arbitrary point at the centre of the gray

square, but these points have no features that will allow the visual system to identify them as corresponding. In contrast, it is much easier to solve the correspondence problem for points at the corners and on the boundaries of the square. What this simple example suggests is that one possible way of solving the correspondence problem is to match up boundaries and edges and then proceed by mapping points internal to bounded objects and surfaces. Although there will be disparities between the two retinal images, there will also be significant regions of overlap and correlation. How might the visual system exploit these regions of overlap and correlation to solve the correspondence problem? Churchland and Sejnowski offer the following hypothesis (1992, pp. 199–202). Imagine a 3-D scene containing a dog in front of a fir tree in front of a barn.

> In trying to break down the problem, the key fact is that portions of right and left retina may be highly correlated, in the sense that patterns of gray levels over a stretch of the left retina will be very similar to a pattern in right retina, but shifted by an amount determined by the relative depth of the perceived object from the plane of fixation. To get the correct conceptual bead on how this fact might be useful, envision the situation by analogy. If, god-like, we could slide the two images past each other in the horizontal plane, we could quickly find a registration between the two dog images in the foreground and, sliding a bit further, one that lines up the fir tree images but not the dog images, and finally, one that lines up the barn but not the fir tree or the dog, though, to be sure, the lining up is only approximate.
>
> (ibid., p. 201)

The key idea is trying to map the two retinal images onto each other (rather than mapping individual points onto each other) at different degrees of depth, on the assumption that as the depth increases so too does the horizontal displacement between the two images. Churchland and Sejnowski use this basic idea to define a compatibility function that will map points in the two retinal images onto each other relative to different degrees of displacement. They discuss a neural network designed by Paul Churchland that can compute the compatibility function relative to a given degree of displacement. The details of how this works are complex, but the important point is that the task (solving the correspondence problem) is characterized in a way that effectively turns it into a pattern recognition task – into a matter of trying to map images on to each other. There is a sense in which this type of pattern recognition task can be described as a process of hypothesis formation and testing. But it is not hypothesis formation and testing in the way that Fodor understands it – and it is far from obvious that it requires anything like a language of thought.

This is an area, of course, in which it would be rash to try to draw any sweeping conclusions. We have only discussed general features of constraint satisfaction and one very basic problem in early visual processing. Nonetheless,

it certainly seems that opponents of the language of thought hypothesis have a number of resources at their disposal to contest the argument from the underdetermination of perceptual information to the conclusion that perceptual integration requires a syntactically specifiable computational medium. This seems to be an area where considerable further work is required both to clarify the general tasks that are being performed and to elucidate the mechanisms that might be carrying out those general tasks.

10.6 Concept learning

Fodor's understanding of the problem of concept learning is driven by experimentation into the categorizing abilities of non-human animals. He takes the problem to be determining the environmental conditions under which a designated response is appropriate. So, for example, we know that pigeons are capable of very sophisticated forms of visual discrimination, such as discriminating colored slides that contain images of people from slides that do not. They can reliably pick out scenes that contain images of a particular individual from slides that do not – and in fact they can distinguish slides with pigeons from slides with other birds.[6] According to Fodor, the pigeons in these experiments are learning to correlate a particular set of environmental conditions (particular images on slides) with a designated response (standardly a pecking response). Learning to perform this correlation is, Fodor thinks, essentially a process of inductive inference.

> What the organism has to do is to extrapolate a generalization (all the positive stimuli are *P*-stimuli) on the basis of some instances that conform to the generalization (the first *n* positive stimuli were *P*-stimuli). The game is, in short, inductive extrapolation, and inductive extrapolation presupposes (a) a source of inductive hypotheses (in the present case, a range of candidate values of *P*) and (b) a confirmation metric such that the probability that the organism will accept (e.g. act upon) a given value of *P* at *t* is some reasonable function of the distribution of entries in the data matrix for trials prior to *t*.
>
> (Fodor 1975, p. 37)

As with perceptual integration, Fodor sees this as essentially a task of hypothesis formation and testing. The pigeon needs to represent the data (namely, the slides that have been rewarded as a subset of the total slides that have been presented) and then formulate a hypothesis about which features of slides are correlated with the reward. This hypothesis is either confirmed or disconfirmed by subsequent episodes. When the learning process is understood in these terms, it is clear why it requires a representational medium as strong as the language of thought.

6 See Walker (1983, pp. 254–266) for a brief survey.

As in the case of perceptual integration, however, there are questions to be asked about how Fodor characterizes the task that is being performed. Fodor's task analysis is very high-level, and he explicitly draws upon accounts from the philosophy of science of how inductive inference works. Fodor's concept learners are feathered scientists. Unsurprisingly there are alternative accounts of what is going on in this type of learning. The first point to note is that we are dealing here with a classic example of *conditioned behavior*. Conditioning takes place when an organism learns to associate a particular response with a particular stimulus. Conditioning is taken by many theorists to be a form of associative learning that works by building up an association between the conditional stimulus (the sound of the bell, in Pavlov's experiments, or the presentation of the slide with the pigeons in it) and the unconditional stimulus (which is the reward). Considerable research has been done on understanding the mechanisms of conditioning, particularly with respect to classical conditioning.[7]

It is potentially very significant for how we think about discrimination learning that none of the currently popular theories of conditioning treat it as a process of hypothesis formation and testing in the way that Fodor suggests. Instead the dominant approach is to treat the process of learning through conditioning in terms of associations being established between representations of events/actions, where those representations are not in any sense sentential. Consider, for example, the generic *associative-cybernetic* model proposed by Dickinson and Balleine as a way of capturing some of the dominant ideas in contemporary thinking about conditioning (Dickinson and Balleine 1993).[8] Here is how they summarize the model:

> The basic idea is that being in a particular situation makes the agent imagine performing a response, but not very vividly. If this response has been previously associated with a goal, the agent will, as a consequence of the associative properties of the model, also think of the goal. The cybernetic component reflects the fact that imagining the goal feeds back to enhance the response image until, in the spirit of the ideo-motor theory, it is sufficiently vivid to trigger the response as an overt action.
>
> (1993, p. 281)

7 Dickinson (1980) is an excellent introduction to the theory of conditioning.

8 Dickinson and Balleine set up the associative-cybernetic model in order to argue that it cannot accommodate all cases of instrumental conditioning and hence that will sometimes need to appeal to some form of belief–desire psychology in making sense of instrumentally conditioned behavior. Two points are worth making, however. First, they clearly accept that the associative-cybernetic model will explain the vast majority of cases of instrumental conditioning. And second, the version of belief–desire psychology that they propose does not involve attributing anything like the sort of processes of hypothesis formation and testing that Fodor discusses. It is far from clear that deploying beliefs and desires to explain the behavior of non-linguistic creatures brings with it commitment to the language of thought hypothesis. For further discussion of how models of psychological explanation can be applied in the non-linguistic domain, see Bermúdez (2003a).

It is clear that the representations in question are imagistic, rather than sentential in nature. In this sense the associative-cybernetic model fits easily with the non-propositional account of non-linguistic thought briefly discussed in section 10.4.

Of course, there are some difficulties in applying the associative-cybernetic model to the pigeon example we are discussing. As Fodor would no doubt be quick to point out, there is more going on here than simply being in a certain situation and imagining a particular response. The pigeon has to learn to make the response when faced with a colored slide *that has the appropriate features*. What needs to be explained is how the pigeon discriminates, for example, between slides that contain pigeons and slides that do not contain pigeons. It is of course here that the process of hypothesis testing is supposed to be required. One might wonder, though, whether identifying similarities between slides should not be understood as an essentially perceptual process. It seems plausible that all organisms capable of learning have built into them some form of similarity metric that will permit the detection of salient similarities. Quine elegantly made the point some time ago:

> If an individual learns at all, differences in degree of similarity must be implicit in his learning pattern. Otherwise any response, if reinforced, would be conditioned equally and indiscriminately to any and every future episode, all these being equally similar. Some implicit standard, however provisional, for ordering our episodes as more or less similar must therefore antedate all learning, and be innate.
>
> (1974, p. 19)

Quine is surely correct that some form of perceived similarity must precede any inductive generalization, if the agent is even to get started on the process of extrapolating from past experience. The hypotheses that the pigeon is supposed to be formulating (on Fodor's account) are presumably hypotheses about which similarity is salient – and hence the formation of hypotheses presupposes the detection of similarities. But then it is natural to wonder whether this basic capacity to perceive similarities is not sufficient to explain the type of learning under discussion.

Fodor would no doubt have reservations about this attempt to assimilate discrimination learning to operant conditioning. He is emphatic that discrimination learning should be taken to underpin operant conditioning, rather than to be identical to it.[9] The process of conditioning works because the pigeons have learnt to discriminate slides with pigeons on them from slides without pigeons on them – rather than vice versa. It is not clear, however, that this point really tells against the suggestion that what is really

9 See, for example, the lengthy footnote 6 on pp. 35–36 of Fodor (1975).

going on is a simple registration of similarities according to an innate metric, as opposed to a semi-formal process of hypothesis-formation and testing. This simply seems to be an issue where considerably more research and analysis is required.

As with practical decision-making and perceptual integration, it seems clear that we are a long way from understanding what is going on in what Fodor terms concept learning. It is, moreover, equally clear that Fodor's arguments that concept learning requires a language of thought are far less compelling than he takes them to be. Before ending this section, however, we should briefly address another argument that Fodor puts forward against the claim that public language is the only language of thought we need. The argument, in essence, is that any appeal to public languages as an alternative to the language of thought is doomed because the very possibility of learning a public language requires a language of thought.

Once again the argument is an argument from the need for hypothesis formation and testing to the language of thought. The central claim is that language learning is essentially a process of hypothesis formation and testing. We learn a language, according to Fodor, by gradually converging onto correct hypotheses about the meanings of words. Linguistic understanding is essentially rule-based. The language of thought comes into the picture as the medium in which those rules are formulated – and hence, of course, as the medium in which hypotheses about the nature of those rules are formulated during the process of language learning. Both understanding a language and learning a language are taken to involve translating sentences of that language into another language, the understanding of which can be taken for granted – namely, the language of thought. Given this, Fodor argues, the very possibility of language learning requires a representational medium at least as expressively powerful as the language being learnt.

The rules are what Fodor terms truth-rules.[10] A truth-rule specifies a referent for a given singular term and an extension for a given predicate (in the case of syncategorematic expressions such as the logical particles the truth-rules will specify introduction and elimination rules). The truth-rule for an arbitrary predicate '– is *F*' will take roughly the following form.

> '*x* is *F*' is true iff *x* is *G* (where an arbitrary singular term can take the place of '*x*' and '*G*' is a predicate co-extensive with '*F*').

The truth-rule for a singular term '*a*' will be along the following lines.

> '*a* is *H*' is true iff *b* is *H* (where '*a*' refers to the same object as '*b*' and an arbitrary predicate can take the place of '*H*').

10 It may well be that there is more to linguistic understanding than mastery of the relevant truth-rules, but Fodor is adamant that mastery of truth-rules will be a *necessary* condition of linguistic understanding.

Fodor's proposal is that giving the truth-rule for a given sentence involves a translation of that sentence into the language of thought.

Enthusiasts for the theory of meaning will recognize that these truth-rules are significantly stronger than the truth-rules envisaged in standard truth-conditional theories of meaning, such as those canvassed by Davidson (Davidson 1967). Truth-conditional theories of meaning exploit disquotational truth-rules in which the same sentence is both *mentioned* (in the left-hand clause) and *used* (in the right-hand clause). A typical truth-rule within a truth-conditional theory of meaning will take the form

'The car is in the garage' is true iff the car is in the garage.

Fodor's truth-rules are clearly not disquotational in this sense. The clause that gives the truth-conditions of the target sentence does not use the words that feature in the sentence – rather, it uses words that are co-referential (in the case of singular terms) and co-extensive (in the case of predicates). One immediate question that arises, therefore, is whether Fodor's way of thinking about linguistic understanding and language learning might not be excessively demanding.

It is certainly the case that most proponents of truth-conditional theories of meaning would take issue with Fodor on this point. Theorists of meaning inspired by Davidson do not feel it necessary to go beyond disquotational truth-rules. On the other hand, however, there is a case to be made for saying that Fodor's project is somewhat different from that engaged in by theorists of meaning. Fodor is interested in explaining what it is to understand a language in a way that will explain what is involved in learning a language. This imposes an explanatory burden over and above what a truth-conditional theory of meaning takes itself to be elucidating. We can put the point by saying that on Fodor's view, a theory of meaning is a theory of understanding, and there is some plausibility in the claim that disquotational truth-rules will not yield a theory of understanding, in at least the following sense. Nobody who does not already understand the sentence "The car is in the garage" will learn anything from being told that

"The car is in the garage" is true iff the car is in the garage.

There is considerable plausibility, then, in Fodor's assumption that a theory of meaning needs to be more "full-blooded" than would be possible using disquotational truth-rules.

Nonetheless, there is room for skepticism about whether a theory of understanding really does need to take the rule-based form that Fodor discusses – and hence, correlatively, about whether we should model language-learning as a process of forming and testing hypotheses in the language of thought about what those rules are. Consider how Fodor's model might

work in the case of color predicates. The truth-rule for the predicate ' – is red' will be along the following lines.

> '*x* is red' is true iff *x* is red* (where *x* can be replaced by an arbitrary singular term and 'red*' is a predicate in the language of thought that picks out all and only red things).

By extension, learning the predicate ' – is red' will be a matter of formulating different versions of the rule with different language of thought predicates on the right-hand side of the rule until the correct formulation involving 'red*' is eventually reached.

It is clear, however, that there are accounts of how we go about learning the meaning of 'red' that do not involving finding some mapping from 'red' to a co-extensive word in the language of thought (or any other language). On the picture that lies behind Fodor's argument, learning the meaning of a term is primarily a matter of learning what falls within its extension – and learning what falls within its extension is a matter of finding a way of specifying that extension in other words (as when I understand the French word 'rouge' by grasping that it picks out all and only the same objects as 'red'). But one might think that this does not do justice to a very basic fact about learning the meaning of color words – a fact deriving from the observational nature of color words. Identifying objects as red is something that we do as a function of seeing them as red, so that learning the meaning of 'red' is a matter of learning to respond to particular types of perceptual experience – of learning that 'red' is the word one uses when one wants to describe the color of things that look red.

Of course, Fodor is making a claim about necessity rather than sufficiency. His argument is that one could not understand the meaning of 'red' without knowing the truth-rule for 'red', and he is certainly not committed to the further claim that knowing the truth-rule for 'red' is sufficient for understanding the meaning of 'red'. Nonetheless, doubts about the sufficiency claim seem to carry over to the necessity claim. If we grant that learning the meaning of 'red' is at least in part a matter of responding appropriately to perceptual experience, then one might reasonably ask for a more detailed account of how this is supposed to work. One very plausible account would be to say that we learn how to apply the word 'red' through learning how to identify when objects are similar in the appropriate sort of way. That is to say, learning the meaning of 'red' is a matter of learning what similarities-with-respect-to-color count as similarities in redness – and understanding the meaning of 'red' is being able to identify when objects are similar in the appropriate ways. On this view the process of learning begins with appreciation of paradigm cases of redness, together with an appreciation of paradigm cases of competing colors, and takes the form of gradually becoming more sensitive to different types of similarity to those competing paradigms. This would sit naturally with the view that a proper understanding of 'red' consists in being able correctly to

identify the appropriate similarities in the appropriate contexts, so that understanding 'red' is fundamentally a matter of having a properly tuned perceptual system and being able to recognize which similarities are salient in particular contexts.

If this is right then, at least in the case of explaining what it is to understand and learn the meaning of 'red', there may be no need for the notion of grasping a truth-rule at all. It may be that learning the meaning of 'red' is a matter of learning how to navigate the similarity space of colors, rather than forming hypotheses about the extension of the word. In fact, the importance of recognizing perceptual similarity in classifying things according to color has led many theorists to the thought that this is precisely the sort of cognitive task that is best modeled by artificial neural networks. As has been stressed at a number of points in this book, neural networks lend themselves particularly well to modeling cognitive tasks involving recognizing patterns and detecting similarities. This seems particularly so in the case of color perception, and hence by extension in the application of color vocabulary. The fundamental problem in understanding how we think and speak about color is understanding how our coarse-grained color concepts and vocabulary are super-imposed on the very fine-grained color discriminations that we are clearly capable of making. A very natural way of thinking about this is in terms of the "pull" of paradigm examples of different concepts. In given contexts an object will seem to be more similar to the paradigm of red than it is to the paradigm of orange – that is to say, the pull of the red paradigm will extend further over the color solid than the pull of the orange paradigm. Artificial neural networks offer a very natural way of modeling this type of context-sensitive similarity judgment.

Of course, there are questions to be raised about just how representative color words are for exploring the general contours of language learning. Unlike color words the vast majority of words that we learn are not *observational*. We can learn what they mean and how to use them without having any first-hand experience of what they name – whether that is an object, as in the case of a singular term, or a set of objects, as in the case of predicates. One might wonder, therefore, whether something along the lines of Fodor's truth-rules is required for understanding words that are non-observational, even if truth-rules are not required for observational concepts.

This might turn out to be the case, but there are many accounts of what concepts are and how we should think about linguistic meaning on which this fails to follow. The approach to concepts and linguistic meanings in terms of their extension has not been very popular among psychologists studying language-learning and concept formation (Prinz 2002). There is also a range of alternative theories in philosophy. Many philosophers have thought, for example, that we should understand meaning in terms of use – that understanding the meaning of a word should be explained in terms of the practical abilities that a speaker manifests in using that word. This line of thought stands in explicit opposition to the idea that meaning is to be

given in terms of truth-rules. Its most famous exponent is Ludwig Wittgenstein, and it has recently been powerfully advocated by Paul Horwich (Horwich 1998), neither of whom think that Fodorean truth-rules have any role to play in linguistic understanding.[11] Michael Dummett has long been an advocate of an approach to understanding linguistic meaning that tries to combine the truth-conditional approach with the principle that meaning is use. Dummett attempts to reconcile the two approaches by arguing that grasping the truth-conditions of a sentence is not a basic capacity, but rather something that itself needs to be explained (Dummett 1973). The direction of explanation that Dummett offers goes via the idea that knowledge of the truth-conditions of a sentence is derived from knowledge of what it would be to go about establishing the truth-value of that sentence. Our understanding of subsentential units of meaning, such as names and predicates, is derived from our understanding of how to go about establishing the truth-value of sentences in which they feature. Once again, there is no appeal to Fodorean truth-rules.

At best, therefore, the discussion is inconclusive. Fodor's argument for the language of thought hypothesis rests a far from obviously compelling account of what it is to understand a language and to learn a language. It may turn out in the long run that Fodor's account is the right one – and hence that some version of the language of thought hypothesis is true. But as things currently stand, Fodor's theory is certainly not "the only game in town" and many would think that it is one of the less plausible views on the market.

11 Horwich favors disquotational truth-rules, of the type discussed above. These are very different from Fodor's truth-rules.

Concluding thoughts

Toward a fifth picture

We have been guided in thinking about the philosophy of psychology by four dominant pictures. Each picture incorporates a different set of metaphors and tools for thinking about the mind and how it relates to the brain and to the environment. Each highlights different aspects of the mind and offers a distinct way of responding to the interface problem. The representational picture is built around the metaphor of the mind as computer, treating cognitive abilities in terms of computational tasks and using the idea of computation as the thread linking together different levels of explanation. According to the functional picture, in contrast, the causal dimension of the mind is paramount. Instead of focusing on particular cognitive abilities the functional picture highlights the causal dimension of individual mental states, using the role/realizer relation to show how what goes on at lower levels of explanation can be causally relevant to the personal-level states of commonsense psychology. While the functional and representational pictures try to tackle the interface problem head-on, the pictures of the autonomous mind and the neurocomputational mind try in their very different ways to undercut its force. The picture of the autonomous mind highlights what it takes to be the uniqueness and irreducibility of personal-level psychology, deriving this uniqueness from the norms of rationality claimed to govern personal-level psychology. The picture of the neurocomputational mind, in contrast, is strongly committed to the metaphor of the mind as brain and accepts that our thinking about the mind must co-evolve with our thinking about the brain in a way that may lead to significant revisions of our commonsense ways of understanding cognition and behavior.

Each picture of the mind emphasizes different aspects of cognition and works on the basis of different paradigms. The neurocomputational picture, for example, stresses what one might think of as low-level cognitive mechanisms. It takes issue with the natural assumption that high-level cognitive achievements must be carried out by complex computational mechanisms. Instead, it emphasizes the explanatory power of surprisingly simple mechanisms performing operations of template-matching and pattern recognition. The plausibility of the neurocomputational view is in large part a function of how convinced one is by neural network models of higher cognitive abilities (and indeed of how representative one takes neural networks to be of neural functioning). The autonomy view, on the other hand, takes as its paradigms of cognition the most sophisticated forms of rational reflection and deliberation. The types of thinking highlighted by the autonomy view

are not simply *governed by* norms, but rather *guided by* norms in ways that involve reflecting on the demands imposed by norms of rationality. The representational and functional pictures fall somewhere between the two. One basic idea behind the representational approach is that formal transitions between syntactic entities can track semantic transitions. This is of interest primarily in connection with types of thinking that lend themselves to being codified in formal models such as expected utility theory or deductive logic. Whereas the representational picture sees thinking in primarily logical terms, the functional picture takes a causal view of the dynamics of thought. The paradigm for the functional picture is the interaction of beliefs and desires in the generation of behavior. Representational theorists take the challenge to be explaining how logical transitions can be captured by causal transitions. Functional theorists, in contrast, take causal transitions between mental states as basic and see the challenge as showing how those causal transitions can be used to characterize the mental states featuring in them.

Each of the four pictures we have been considering adopts a broadly similar strategy. This is the strategy of trying to show that the mind as a whole should be understood on the model of the favored paradigm types of thinking. It is predictable where the difficulties will be found. One might reasonably think, for example, that the neurocomputational approach will have difficulties with the deductive transitions and probabilistic calculations taken as paradigmatic by proponents of the representational mind. It is true that theorists probably underestimate the extent to which logical reasoning is a matter of pattern recognition – after all, one can only apply formal rules if one can identify which formal rule is salient in a particular context, and this is often a matter of seeing what pattern is exemplified by a given inference. But it seems likely that the rule-governed nature of logical reasoning will make it difficult to capture with the resources of the neurocomputational approach. By parity of reasoning one might expect the perceptual and recognitional abilities highlighted by the neurocomputational approach to pose problems for representational theorists. Even though perceptual processes are no doubt governed by rules, these rules seem fundamentally different from the inflexible and formal logical rules that are easily captured and manipulated in the language of thought. It is certainly true that researchers in traditional artificial intelligence (what is sometimes called "good old-fashioned artificial intelligence") have had far more success in modeling formal and semi-formal types of cognition that they have had in developing models of perceptual processing.

Similar difficulties arise with the different emphases and priorities of functional and autonomy theorists. Surely, autonomy theorists will ask, there must be more to theoretical deliberation and practical reasoning than causal interactions between mental states. How can a purely causal story do justice to our more reflexive and reflective modes of thinking? And of course the same problem arises in the other direction. The rarified approach proposed by autonomy theorists seems to involve too much heavy-duty

machinery to provide a plausible account of the myriad of trivial inferences and uncomplicated predictions that make up daily psychological life. How much time do we really spend thinking about "how things ought to be", as opposed to making quick and efficient guesses about "how things are"?

It has not gone unnoticed that the general approaches to the mind we have been considering each work best for a limited domain. One obvious response is to try to show that thinking and cognition are really far less varied than they initially appear. So, for example, a neurocomputational theorist might attempt to show that cognition is far less rule-governed and language-dependent than it initially appears to be, while a functional theorist might try to show that the norms governing practical reasoning and deliberation can be understood in causal terms. Another response would be to try to finesse the situation by locating different approaches at different levels of explanation. As we observed in Chapter 9, it is standard for supporters of the representational approach to argue that it is not directly in competition with the neurocomputational approach, because the neurocomputational approach is best viewed as an account pitched at the implementational level. Similarly, autonomy theorists frequently argue that the causal approach adopted by functional theorists is best seen as an account of the subpersonal underpinnings of cognition, rather than of personal-level thought.

It seems unlikely, however, that the strategy of either assimilating the competition or trying to show that there is no real conflict by locating the apparent competition at a different level of explanation will prove completely satisfying. Thinking and cognition are just too complex and variegated. In the light of this it is natural to wonder whether trying to find a single monolithic account of the mind as a whole is really the best strategy. Perhaps it would be more profitable to explore the possibility of combining some of the insights and analyses offered by the different approaches. In the remainder of this concluding chapter I would like to make some very preliminary and programmatic remarks about one possible way of developing such an alternative account. What follows draws upon some of the arguments and claims that have emerged in the main body of the book, but is very much a personal view. The suggestions that follow represent one way of navigating through the complex issues in this area, but it is certainly not the only way, and there may well be better ways.

Let me begin by drawing attention to some ideas that have come to the surface in the course of this book. One theme that has emerged at various points has to do with the significance of commonsense psychology. All four pictures of the mind we have been examining take commonsense psychology to play a fundamental role in our understanding of ourselves and others – so much so that we were able to characterize the four pictures in terms of their different responses to the problem of explaining how the explanatory framework of commonsense psychology interfaces with explanatory frameworks lower down in the hierarchy of explanation. Commonsense psychology is an explanatory tool that explains and makes sense of behavior by interpreting it

as the result of beliefs, desires and other propositional attitudes. A commitment to the explanatory power of folk psychology fits naturally with the view that beliefs, desires and other propositional attitudes are the "springs of action". The simplest explanation of the explanatory success of commonsense psychological explanations is that they work because they are true, which is to say that they work because they correctly identify the beliefs and desires that really caused the actions in question. And similarly for prediction. One might think, therefore, that whenever we are dealing with behavior that cannot be seen as a direct response to some environmental stimulus we must be dealing with action that is in some sense generated by propositional attitudes. As we saw in Chapter 8, this way of thinking about the springs of action brings with it a particular interpretation of the architecture of cognition – specifically, a sharp distinction between "central" cognitive processes that involve propositional attitudes and "peripheral" cognitive processes that are not defined over propositional attitudes but instead provide inputs to the propositional attitude system. These modular processes have certain characteristics (such as informational encapsulation, domain-specificity, speed, and so on) that make it natural to classify them as subpersonal, in opposition to the personal-level propositional attitude system, which has none of these characteristics.

We have seen a number of ways of putting pressure on this way of thinking about the architecture of cognition. In Chapter 7 we looked at ways of making sense of the behavior of others that do not involve the attribution of propositional attitudes and hence that do not involve the explanatory framework of commonsense psychology. Much of our understanding of other people rests upon a range of relatively simple mechanisms and heuristics that allow us to identify patterns in other people's behavior and to respond appropriately to the patterns detected. The simplest such patterns are a function of mood and emotional state, while the more complex ones involve social roles and routine social interactions. One interesting feature of these modes of social understanding is that, by downplaying the role of the propositional attitudes in social understanding, they diminish the centrality of the interface problem in our thinking about the mind. These are personal-level modes of social understanding that do not bring with them the complicated theoretical machinery that philosophers of psychology have standardly taken to be required for navigating the social world. They do not require maneuvering oneself into another person's perspective on the world (in the manner proposed by the simulationist approach to social understanding), or bringing to bear a tacitly known theory of cognition and behavior (as suggested by theory-theorists).

Of course, our ways of explaining behavior are not invariably a good guide to how that behavior came about. Optimal foraging theory is a striking example, where a complex theoretical framework is used to explain and predict behavior generated by a set of very basic mechanisms and rules. But the discussion of ways of thinking about the path from perception to action

in Chapter 8 suggested that there is a range of ways of generating behavior that are neither reflex or instinctual, nor are mediated by propositional attitudes. One important idea that emerged from that chapter is that the line between perception and cognition may not be as sharply defined as it is standardly taken to be. There are ways of perceiving the world that have direct implications for action. Frequently what we perceive are the possibilities that the environment "affords" for action, so that we can act on how we perceive the world to be, without having to form or exploit beliefs and other propositional attitudes. Admittedly, the perception of affordances is a phenomenon at the personal level of explanation, and one should be wary of drawing conclusions about the structure of subpersonal cognitive architecture from facts about the nature of personal-level thought. But the perception of affordances cuts across the sharp distinction between, on the one hand, peripheral, domain-specific and informationally encapsulated modules providing a "neutral" representation of the distal environment and, on the other, central cognitive processes defined over the propositional attitudes.

The discussion of the massive modularity hypothesis in Chapter 8 put further pressure on the standard distinction between peripheral and central processes. According to the massive modularity hypothesis, there is no such thing as domain-general thinking. All thinking is subserved by domain-specific modules that evolved to deal with specific problems confronted by our hominid or primate ancestors. These so-called Darwinian modules are very different from the modules discussed by Fodor. They are not informationally encapsulated, for example, and their principal function is not to transform sensory input into a format that can serve as input into central processing. They are modular in two senses. First, they are domain-specific – engaging only in response to a limited set of inputs and applying only a limited set of operations to those inputs. Second, the representations they employ are not best viewed in terms of the categories of propositional attitude psychology.

How should we respond to these pressures on the standard distinction between subpersonal modular processing and a personal-level propositional attitude system? One response would be eliminativism about the propositional attitudes, effectively holding that the propositional attitudes should have no role to play in how we think about the genesis of behavior – and hence, *a fortiori*, no role to play in social understanding. Such an approach would mesh well with some ways of developing the neurocomputational approach to the mind – in particular with the views put forward by the Churchlands. On the other hand, however, one might wonder whether eliminativism is too drastic a response. Perhaps it would be better to circumscribe the role of the propositional attitudes, rather than to banish them altogether. The most obvious way of doing this would be to break the connection between intelligent behavior and the propositional attitudes by accepting that there are many ways of behaving in a non-instinctual and non-reflex manner that completely bypass the propositional attitudes. These

are forms of behavior that we can explain and understand quickly and efficiently without bringing to bear the machinery of propositional attitude psychology.

Of these two possible responses, the balance of the arguments in the main body of the book seems clearly to point to the second, less drastic response. It is hard to imagine that all our talk of propositional attitudes will turn out to have been completely mistaken and that all the work that we take to be done by the propositional attitudes will turn out to be performed by Darwinian modules, mechanisms of template-matching and pattern-recognition, and ways of accommodating oneself to established social routines. It is more plausible to think that the propositional attitudes do have a very real role to play in certain types of thinking and in the genesis of certain types of behavior – particularly where we find the types of norm-guided thinking highlighted by autonomy theorists and the logical thinking emphasized in some of the arguments for the language of thought hypothesis.

One might try to accommodate these various pressures at the level of cognitive architecture by revising the standard distinction between central and peripheral processing in favor of a three-way picture distinguishing two fundamentally different forms of personal-level cognition, in addition to the peripheral modules responsible for processing sensory input. Personal-level cognition can involve either the complex processes and mechanisms defined over the propositional attitudes or the much simpler Darwinian modules, heuristics, and mechanisms of template-matching and pattern recognition that we have been discussing. The suggestion here is not that we interpose an additional set of mechanisms between peripheral modules and central cognition, but rather that we think of there being two fundamentally different personal-level routes to action, one engaging the propositional attitudes and the other engaging evolutionarily more primitive mechanisms that are faster and more specialized. The standard distinction between peripheral processing and modular processing can be visualized two-dimensionally, as a core of central processing bounded by an input layer and an output layer of peripheral modules. The current view is best construed in three-dimensional terms, with the propositional attitude system superimposed upon a complex network of pathways leading from peripheral input modules to peripheral output modules. Some of these pathways correspond to Darwinian modules and others to heuristics and social routines. Each pathway leads from input modules to output modules without engaging the propositional attitude system. We might think of each individual pathway as working to solve a particular set of problems in response to a particular type of input. It may be, for example, that one of these pathways corresponds to the so-called cheater detection module, processing inputs of social situations to search for free-riders. On the view being suggested, the cheater detection pathway does not work to produce beliefs – it does not feed directly into the propositional system. Rather, it has immediate implications for action. The problems it solves are problems of how to behave in particular situations. These are

problems, crudely speaking, of whether or not to cooperate, with the question of what is to count as cooperation clearly fixed by the context in which the issue arises. Once the cheater detection module has done its work there is standardly no need for further processes of practical reasoning involving the propositional attitude system – although of course there are different ways of reacting to the presence of a free-rider and there has to be *some* way of deciding between them.

Three significant challenges naturally arise at this point. The first has been briefly considered in section 8.4 in the context of Fodor's argument against the massive modularity hypothesis. As Fodor points out (Fodor 2000), there is a lack of fit between the outputs of peripheral modules (what we might think of as Fodorean modules) and inputs to Darwinian modules. As we have seen at various points in the book, we should think of the Fodorian modules that collectively comprise the early visual system as collaborating to produce a representation of the three-dimensional layout of the distal environment that has only a rudimentary degree of interpretation. The cheater detection module, however, requires highly interpreted inputs. It will only work on representations of social exchanges – and indeed only on those social exchanges that have a cost–benefit dimension. Clearly there needs to be some further processing intervening between the end of peripheral processing and the various pathways that we have been discussing. The first issue, then, is giving an account of this processing and how it fits into the overall architecture of cognition. This is not a topic that has received any attention in the psychological or philosophical literature. We are dealing with processing that effects a form of filtering, working to parse and interpret the deliverances of the modular sensory systems into a format that will engage one or other of the Darwinian modules or other pathways from perception to action. As such, it will be a form of domain-general processing. However, as we saw in section 8.4, there is no need to follow Fodor in the claim that it will have to engage what he thinks of as the domain-general propositional attitude system. A proper development of the position being sketched out here will need to offer a substantive account of this type of intermediate domain-general processing. It is very possible that research into artificial neural networks will be illuminating in this area. The filtering tasks that need to be carried out at this level may well turn out to involve the type of detection of patterns and sensitivity to prototypes that artificial neural networks are so good at modeling.

We can view the first challenge as demanding an explanation of how a particular form of selection problem is solved. This is the selection problem of determining which of the various possible perception–action pathways should be engaged in a particular context. But this is not the only selection problem that needs to be solved. I have suggested that processes and mechanisms involving propositional attitudes are superimposed upon the more primitive framework of perception–action pathways. But what determines whether and when these processes and mechanisms are engaged?

Again, we are not in a position to make anything more than some very general comments. We can view the propositional attitude *complex* (a better terminology, I think, than the widespread talk of the propositional attitude system) as coming into play to deal with situations that cannot be dealt with by the lower-level perception–action pathways. This would occur most obviously when we are dealing with types of thinking that are not a response to particular demands imposed by the immediate environment – forms of reflection, deliberation and forward planning that are not stimulus-driven. It is no accident that these are taken as paradigmatic types of thinking by those who see the propositional attitudes as central to cognition. But one might also expect elements of the propositional attitude complex to be engaged in the face of stimuli that do not fall neatly into the domain of one and only one perception–action pathway. It may not be possible to parse certain unfamiliar situations into a format that will serve as input into one or other pathway. In such a situation one might expect that background beliefs will need to be brought into play. Conversely, as we saw when discussing the massive modularity hypothesis in section 8.4, there may be situations that fall within the domain of more than one perception–action pathway – a situation, for example, that comes within the ambit both of the cheater detection pathway and the danger avoidance pathway. In such circumstances the two pathways may come up with different and incompatible actions. The resources of the propositional attitude complex may be required to resolve the conflict. But how does this take place? How are conflicts between perception–action pathways identified? How are unfamiliar situations "handed over" to the propositional attitude complex? These are all questions that call for considerable further study.

The third challenge in this area is to give a principled account of the significance of natural language in cognition – and in particular of the relation between natural language and the propositional attitudes. This is important if we are properly to evaluate the various arguments for the language of thought hypothesis considered in Chapters 9 and 10. The force of those arguments was that the propositional attitude complex must be explained independently of natural language, because we can only give an account of what it is to learn and understand a natural language in terms (*inter alia*) of beliefs about the means of words – beliefs that cannot themselves be in any sense dependent upon natural language. We considered an alternative to the language of thought hypothesis. This is what I termed the rewiring hypothesis, according to which the architecture of cognition is fundamentally changed by the acquisition of language. Learning a natural language makes available a linguistic medium for thinking that can do much of the work that it is claimed can only be done by the language of thought hypothesis, such as for example explaining the apparent systematicity and productivity of thought. The dialectic between the language of thought hypothesis and the rewiring hypothesis is complex, but we can use the proposals about cognitive architecture made above to get them into focus.

It seems clear that the types of information processing carried out by Fodorean modules have nothing to do with language mastery, except for those directly implicated in language comprehension and production. And let us assume (as seems plausible) that perception–action pathways of the type we have been discussing are equally independent of language. This allows us to formulate what is at issue between the language of thought and the rewiring hypotheses as follows. The rewiring hypothesis is committed to two claims. The first is that we can explain what is going on in peripheral modular processing and perception–action pathways without needing to postulate a language of thought. Modular processing and perception–action pathways may well involve the processing of information, but not in a manner that requires a language of thought. The arguments we considered in Chapter 10 trying to show that the language of thought is implicated in basic perceptual processing are obviously very much to the point here. The rewiring hypothesis will stand or fall with the failure or success of those arguments. The tenability of the rewiring hypothesis depends upon being able to develop plausible models of these types of information processing in terms of mechanisms of pattern recognition and template-matching – as opposed, for example, to the mechanisms of hypothesis formation and testing favored by proponents of the language of thought hypothesis. It is certainly too early to come to any firm conclusions about where the balance of the arguments lies, but let us grant the rewiring hypothesis that there is a plausible story to be told in this area. The next question that arises is whether we can explain what it is to learn and understand a natural language in terms of the same type of mechanisms as are involved in modular processing and perception–action pathways. Here matters are even less clear than they are with respect to modular processing and perception–action pathways.

Very little is known about how languages are learnt and understood. Proponents of the language of thought hypothesis have an *a priori* argument aiming to show that languages can only be learnt through processes of hypothesis formation and testing that require a language of thought – and, moreover, that understanding the meaning of words needs to be modeled in terms of meaning rules formulated in a language other than the language being understood. Against this proponents of the rewiring hypothesis can muster a range of empirical considerations and theoretical arguments. As we saw in section 10.6 there is a range of models of linguistic understanding that do not appeal to meaning-rules of the type envisaged by Fodor, and fairly strong grounds for thinking that the meaning-rules approach cannot work for at least some central cases. In section 5.3 we looked at interesting evidence that artificial neural networks trained to perform language-learning tasks reproduce certain of the learning effects discovered in young children.

It is worth drawing attention to some of the theoretical possibilities opened up by the rewiring hypothesis. The most striking is the possibility of explaining the phenomenon of language in complete independence of the propositional attitude complex. This would allow us to appeal to language

in giving an account of the propositional attitude complex. We might think about the vehicles of propositional attitudes in terms of the rewiring of the brain that occurs when language is acquired – as opposed, for example, to thinking of them in terms of physical realizers of functional roles, or sentences in the language of thought. This would open up the way for a version of what in Chapter 5 we described as the co-evolutionary research paradigm. Our thinking about the vehicles of propositional attitudes would co-evolve with discoveries about the changes that take place in neural structure and neural functioning as language develops. This is as yet fairly uncharted territory. Neuroscientists and empirical psychologists have devoted considerable attention to studying the localization of language in the brain, using evidence from lesions and from imaging studies (Garrett 2003). But this research has tended to be insufficiently fine-grained to help with the problems with which we are concerned. The hypothesis is pitched at the level of individual representations – a level at which the appropriate unit of analysis is the small-scale neural population, rather than the functional area. Moreover, the rewiring hypothesis is more concerned with the representational changes that take place within the brain as whole as a consequence of language acquisition – changes that are hypothesized to occur even in areas that are not dedicated to one or other aspect of language processing.

It certainly seems plausible that the ontogenesis of the human infant involves a process of representational change in which types of mental representation of increasing complexity and sophistication become available – and indeed that a comparable process of representational change occurred in human phylogeny. Models of the process of representational change in human infancy have been offered by a number of authors, including Annette Karmiloff-Smith (1992) and Jean Mandler (1992). According to Karmiloff-Smith, the progression towards language acquisition in infancy is marked by a series of representational redescriptions in each of which information becomes more explicit and available to be exploited in a greater number of transitions and transformations. Unsurprisingly, the emergence of language is responsible for the most far-reaching representational redescription. According to Karmiloff-Smith, information becomes fully explicit and available for general use within the cognitive system when it is re-encoded in an essentially linguistic medium. Similar themes occur in a number of models of the evolution of hominid cognition. As we saw briefly in section 10.2, authors such as Merlin Donald and Steven Mithen have suggested that the emergence of language makes possible the integration of different bodies of domain-specific knowledge (Donald 1991; Mithen 1996).

If such accounts are on the right lines, then we have a promising way of approaching the rewiring hypothesis. However, none of the authors mentioned has proposed a detailed account of the possible neural correlates of representational change. Such accounts as exist have emerged from neurobiologists. The selectionist approach, pioneered by Changeux (1985) and developed by Edelman (1989), postulates a "Darwinian" process whereby an

original multiplicity of representational units (groups of synapses for Changeux, neural circuits for Edelman) is selectively pruned, in response to either/both sensory input and intrinsic factors. Another possibility in this area is that representational change is subserved by a process of parcellation (Ebbesson 1984), whereby selective loss of synapses and dendrites leads to increasing differentiation of the brain into separate processing streams. A proper development of the rewiring hypothesis will very likely require building bridges between the neurobiology of representation and more high-level ways of thinking about the nature of representation and the role of representations in cognition.

It is likely, moreover, that a proper working out of the rewiring hypothesis will involve taking seriously the idea that certain types of thinking are actually carried out in a natural language medium. We saw in section 10.1 that there are considerable difficulties with the idea (what I termed the inner speech hypothesis) that all thinking involves the manipulation of natural language sentences. Nonetheless, as emerged in section 10.2, there are certain types of thinking that arguably require a natural language vehicle. Andy Clark has suggested that natural language is the medium for what he calls *second-order cognitive dynamics*, namely, types of thinking that involve explicitly reflecting on one's own cognitive practices, as when one evaluates the reasoning by which one arrived at a particular conclusion, or explores whether a hypothesis is well supported by the available evidence. I myself have extended this suggestion to argue that a natural language medium is required for all types of thinking that have a *metarepresentational* component, that is to say, all types of thinking that involve thinking about thinking (Bermúdez 2003a). Metarepresentational thinking includes what Andy Clark calls second-order cognitive dynamics but extends beyond it to include, for example, thinking that involves ascribing mental states to others (which involves thinking about a thought as the content of another's mental state); that involves conceptions of necessity/possibility and tense (since such notions are best viewed as operators applying to thoughts); and indeed to all types of thinking that involve logic (since logical thought involves reflecting upon the structure and truth-value of thoughts).

The basic argument for the dependence of metarepresentational thinking upon language is that it requires the target thoughts to have vehicles that will allow them to be taken as the objects of thought. Since the paradigm cases of metarepresentational thinking are instances of conscious thinking, these vehicles must be available to conscious thinking. They must, moreover, be vehicles that make the structure of the target thoughts available. This is clearly required, for example, if one is to reflect upon the inferential relations between thoughts.[1] Natural language sentences appear to be the

1 Strictly speaking, this requirement holds only for those inferences that exploit the internal structure of a thought – the type of inferences that are the subject of the predicate calculus. Inferences of the type codified in the propositional calculus depend solely upon the truth-values of the relevant thoughts.

only candidates that satisfy both requirements. Other candidates satisfy one requirement, but not the other. Imagistic representations, for example, are consciously accessible, but do not make the structure of a thought available. Formulae in the language of thought, conversely, make structure available, but are not consciously accessible.

This suggestion about the nature of metarepresentational thinking gives us a further perspective on the project of trying to explain the propositional attitude complex in terms of language. It allows us to see the proposed explanation as having two parts, one focusing on first-order propositional attitudes (those propositional attitudes directed at the world, rather than at one's own thoughts or those of other people). It is to these that the rewiring hypothesis primarily applies. The explanatory task here is to understand how the acquisition of language changes the neural circuitry in a manner that creates potential vehicles for propositional attitudes. The second part of the explanation, in contrast, focuses on second-order propositional attitudes (those involved in metarepresentational thinking). What we are interested in here is showing how these types of thinking involve the explicit manipulation of natural language sentences. In particular, we need to understand the process of manipulating natural language sentences in a way that avoids the problems confronted by the inner speech hypothesis.

Of course, in sketching out the principal claims of this fifth picture of the mind I have concentrated on the benefits rather than the costs. And there are a number of significant outstanding problems that will need to be resolved before the prospects can be viewed in as rosy a light as I have presented them. Some of these we have already discussed – such as the problem of giving a non-metaphorical account of what it is to manipulate a natural language sentence in thought, and the problem of turning the rewiring hypothesis into a substantive theory of the vehicles of first-order propositional attitudes. There is a further problem directly related to an important strand in the arguments for and against the language of thought hypothesis discussed in Chapters 9 and 10. The proposal here is effectively to understand "central" cognition in terms of natural language. Any such proposal has to answer the obvious challenge of explaining what is going on in apparent cases of "central" cognition in creatures that do not possess a natural language. It is well known that cognitive ethologists, developmental psychologists and cognitive archeologists use the language of propositional attitude psychology to characterize the cognitive abilities of non-linguistic and infralinguistic creatures and to explain their behavior in both natural and experimental settings. How should we deal with talk of animal beliefs, or infant knowledge? Here we seem to have examples of propositional attitudes that cannot be understood in terms of language and hence that do not fit the proposed model.

One obvious way of dealing with this potential difficulty would be through the minimalist strategy of refusing to take at face value the explanatory practices of cognitive ethology, developmental psychology and cognitive

archeology. Talk of animals having beliefs about conspecifics or infants possessing bodies of knowledge about objects and how they behave should be taken as shorthand for a more complex explanation in terms of the simpler forms of central cognition that we have been discussing. When developmental psychologists analyze experiments using the dishabituation paradigm by attributing to 5-month-old infants "knowledge" of the principle that objects move on single connected paths through space–time this should be understood as saying that infants are capable of detecting certain patterns in the behavior of material objects and being surprised by material objects behaving in ways that do not conform to those patterns. Similarly, when ethologists claim that certain species of shore birds set out to "deceive" potential predators by "pretending" to be injured, this should be taken as shorthand for a more complex description of their behavior that can ultimately be understood in terms of innate releasing mechanisms or other, more sophisticated perception–action pathways. Some authors have argued that this type of approach is fundamentally mistaken, on the grounds that we have no better perspective than our actual scientific practices for determining the legitimacy of propositional attitude ascriptions (Kornblith 2002). This may be too extreme, but there is some plausibility in the view that, although one might argue about individual cases, the practice of appealing to propositional attitudes in making sense of the behavior of non-linguistic creatures is too well-entrenched to be dispensed with completely.

Nonetheless, rejection of the minimalist strategy would not leave the defender of the language-based approach to explaining the propositional attitudes entirely without resources. One possible approach would be to exploit the distinction between different types of content that is gaining increasing acceptance. A number of philosophers of mind distinguish between the *conceptual content* characteristic of beliefs and other propositional attitudes, and various types of *nonconceptual content* (see the papers in Gunther 2003). Nonconceptual contents share certain fundamental characteristics with propositional attitude contents. In particular, they can be linguistically expressed by means of "that"– clauses and have a degree of structure that marks them off from perceptual and other imagistic states. What makes them *non*conceptual is that they lack certain fundamental features of propositional attitude contents (with the guiding assumption here being that propositional attitude contents are typically composed of concepts). Most authors who appeal to nonconceptual contents hold that they lack the generativity and productivity generally taken to be characteristic of propositional attitude contents. Since one might well think that generativity and productivity are closely connected with domain-generality, and given that that there is some plausibility (as we saw in section 10.2) in the view that non-linguistic cognition lacks domain-generality, it may well be that we need to characterize the content of the propositional attitudes of non-linguistic creatures in nonconceptual terms. It is natural to combine this with the further thought that we should reserve our propositional attitude vocabulary for

states with conceptual content and instead talk of proto-beliefs and proto-desires at the non-linguistic level. Of course, applying the conceptual/non-conceptual distinction in this way would still leave us with the substantive task of making sense of proto-beliefs and proto-desires, but it would allow us to retain the project of explaining the propositional attitude system in terms of language. Nor, one might think, would this be arbitrary or *ad hoc*. The manifest differences between linguistic and non-linguistic cognition make it implausible to think that there is a single category of propositional attitudes that spans both the linguistic and non-linguistic domains.

The possibility is opening up of a picture of the mind completely different from those we have been considering. In place of the standard distinction between input/output modules and a central propositional attitude system, this new picture sees "central" processing in terms of a language-based propositional attitude complex superimposed upon an intricate network of perception–action pathways. The transitions from and to the modular systems on the periphery are effected by systems of domain-general processing that filter the products of modular processing and engage the appropriate perception–action pathways – or, indeed, the propositional attitude system. These filtering systems may well turn out to involve pattern recognition and template-matching of the sort carried out by artificial neural networks. Within the propositional attitude complex we can distinguish two fundamentally different types of cognition. One type of cognition involves first-order, world-directed propositional attitudes and is to be understood at the neural level indirectly in terms of language – that is, in terms of the rewiring that takes place as a function of language acquisition. The second type of cognition involves second-order propositional attitudes, which involve either thinking about thoughts directly, or thinking about the world in a way that requires thinking about thoughts. These are to be understood directly in terms of language, on the assumption that we think about thoughts through thinking about the sentences that express them.

If this picture is viable, then it may well be that we are much closer to understanding the mind than we imagine – or, at least, that we are much closer to having the tools to understand the mind than we imagine. Following on from Marr's pioneering analysis of the early visual system, we have a number of powerful models of modular processing, many of which involve the rapidly expanding resources of computational neuroscience (Churchland and Sejnowski 1992; Eliasmith and Anderson 2003). We also have, in the language of thought hypothesis, an alternative, but nonetheless powerful, theoretical tool for thinking about modular cognition (although, as we have seen, the suggestion that modular processing is a matter of hypothesis formation and testing is far from uncontroversial). It is, moreover, to modular processing that most of the techniques we currently have for studying the brain have been directed. We are moving towards an understanding of the large-scale functional architecture for various types of modular processing, and single-neuron studies have given us some understanding of

what is going on at the level of individual neurons. It is true that, once we move beyond modular processing, techniques for directly studying the brain become less relevant. But the rapidly expanding field of research into artificial neural networks offers great promise for understanding the processing required to interpret and filter the products of peripheral modules. Artificial neural networks may also help us to understand what is going on in the various perception–action pathways that we have been considering. As we move "upwards" to the propositional attitude system, the proposal to understand propositional attitudes through the lens of language allows us to apply our understanding of language and language acquisition to try to make sense of the mechanisms of cognition. The benefits are clearest in the case of second-order propositional attitudes, since the proposal is to understand these directly in linguistic terms. It is true that we have as yet very little understanding of how to think through the general implications of language acquisition for neural circuits not specialized for language. Yet the rewiring hypothesis at least offers a way of bringing together what we know (and are continuing to discover) about language in linguistics, philosophy and the various branches of scientific psychology and using it to inform the study of neural circuits and neural change in neurobiology.

Whatever the fate of the potential approach sketched out in the last few paragraphs, it seems clear that the future of the study of the mind/brain is interdisciplinary. The philosophy of psychology is not just a branch of philosophy that takes psychology and the behavioral and cognitive sciences as its object. It is itself an essential part of the interdisciplinary endeavor of trying to make sense of a highly complex phenomenon that can be studied from a vast range of perspectives. As with all multi- and interdisciplinary endeavors, there is an urgent need for a framework that fits together the different perspectives and levels of explanation. It is here that we find the distinctive contribution of the philosophy of psychology – tracing key concepts through different levels of explanation and trying to develop and think through pictures of the mind that tie together the conclusions and techniques of radically different explanatory projects. These are exciting times and, to borrow the words of a well-known philosopher, it is good to know that we are unlikely to run out of work.

Annotated bibliography

General texts

Reference works

In addition to standard philosophical resources such as the *Routledge Encyclopedia of Philosophy* and the online *Stanford Encyclopedia of Philosophy*, readers of this book will find the *MIT Encyclopedia of the Cognitive Sciences* (Wilson and Keil 1999) and the *Encyclopedia of Cognitive Science* (Nadel 2003) both very useful. Both cover a range of philosophical topics in addition to providing useful introductions to key areas in scientific psychology, cognitive science and neuroscience. The *MIT Encyclopedia* has shorter entries and will be more useful for initial orientation, while the articles in the *Encyclopedia of Cognitive Science* are more in-depth. The *Companion to the Philosophy of Mind* (Guttenplan 1994) is getting a little out of date, but has some useful entries and contains an interesting extended introductory essay. The *Companion to Cognitive Science* (Bechtel and Graham 1998) contains 60 survey articles covering a wide range of topics in cognitive science at a fairly basic level, together with a very useful historical introduction. *The Handbook of Brain Theory and Neural Networks* (Arbib 2003) and *The New Cognitive Neurosciences* (Gazzaniga 2000) are authoritative collections of survey articles.

Introductory texts

There are a number of textbook introductions to the philosophy of psychology. These generally overlap only in part with the current book. It is well worth looking at some other texts to get a sense of how different authors have approached the field. Somewhat dated, but still very worthwhile is Sterelny (1990). More recent texts include Botterill and Carruthers (1999) and Rey (1997), which offers a fairly partisan development of the language of thought hypothesis.

Readers who do not have a background in mainstream philosophy of mind are strongly encouraged to look at a text such as Kim (1998) in order to get a feel for the metaphysical questions that lie behind some of the pictures of the mind we have been considering. Churchland (1988) is old but well worth reading. Other introductory texts include Heil (1998) and Lowe (2000).

There are a number of interesting texts exploring the computational paradigm for thinking about the mind. Crane (2003) is philosophically motivated, while Harnish (2002) provides useful background on the evolution of

cognitive science and on competing computational paradigms. Thagard (1996) is a good, elementary introduction to cognitive science. Dawson (1998) is rewarding but more advanced. In Posner and Raichle (1999) two leading scientists introduce the techniques and results of cognitive neuroscience. Two books by Andy Clark, *Being There* (MIT, 1997) and *Mindware* (OUP, 2001a), provide accessible introductions to recent philosophically relevant work in cognitive science and artificial intelligence. Goldman (1993a) is an engaging introduction to the interface between cognitive science and philosophy.

Anthologies and collections of papers

This volume is accompanied by a collection of readings designed to complement the principal themes (Bermúdez and Macpherson 2005). But of course many important and influential papers have not been included and can be found in other collections. Block's two-volume *Readings in Philosophy of Psychology* (Block 1980) was very influential and is still worth seeking out. Both editions of W. Lycan's *Mind and Cognition* (1990 and 1999) contain much of interest. Cummins and Cummins (2000) is a good collection of papers in the philosophical foundations of cognitive science. Haugeland (1997) contains papers covering the principal contemporary approaches to cognitive architecture and is highly recommended. Goldman (1993b) includes a number of very worthwhile papers. Macdonald and Macdonald (1995a and 1995b) are more specialized, covering more particular debates in a comprehensive manner.

There are numerous collections of papers in "mainstream" philosophy of mind. The most comprehensive are probably Rosenthal (1991) and Chalmers (2002). Heil (2004) contains substantial editorial material. O'Connor and Robb (2003) is worthwhile but less comprehensive. Unfortunately there is substantial overlap across these anthologies. A number of topics relevant to this volume are covered in the specially commissioned essays in Stich and Warfield (2003). The interface between philosophy and the neurosciences is explored in Bechtel *et al.* (2001).

There are a number of useful collections of papers in cognitive science. The most comprehensive is the four-volume *An Invitation to Cognitive Science* (various editors, published by MIT Press in 1995). Posner (1989) is also widely used. E. Lepore and Z. Pylyshyn (eds), *What Is Cognitive Science?* (Blackwell, 1999) contains a number of up-to-date tutorial papers in key areas of the discipline.

Chapter 1

Clarke (2003) is a fascinating account of Descartes's thinking about the mind that does full justice to Descartes's scientific concerns and motivations. There is a good account of Berkeley's theory of vision in Chapters 2–4 of

Pitcher (1977). Kant's and Helmholtz's respective theories of spatial perception are illuminatingly discussed in Hatfield (1990). The picture of Kant as a proto-cognitive scientist is developed in Kitcher (1990) and Brook (1994).

Gardner (1985) provides an interesting historical perspective on the historical emergence of psychology and cognitive science. The same story is told in the introductory essay in Bechtel and Graham (1998). Part I of Harnish (2002) contains a briefer historical introduction to the foundations of cognitive science. Leahey (1992) is an authoritative guide to the history of psychology (see also Flanagan 1984). Finger (1994) is an authoritative history of neuroscience, while Finger (2000) provides an engaging introduction to the development of neuroscience through profiles of a number of key innovators.

The classic application of conceptual analysis in the philosophy of mind and psychology is Ryle (1949). The historical source is Ludwig Wittgenstein, whose views on the philosophy of psychology are illuminatingly presented in Budd (1989). Jackson (1998) presents a sophisticated modern defence of the role of conceptual analysis in a number of different areas of philosophy. See also Chalmers and Jackson (2001), which defends conceptual analysis against the criticisms of Block and Stalnaker (1999).

Putnam's account of law-cluster concepts is presented in Putnam (1962). Many of the issues in the debate over the analytic/synthetic distinction are surveyed in Boghossian (1997). Quine's original article (Quine 1951) has been much reprinted and much discussed. Classic discussions include Grice and Strawson (1956) and Putnam (1965/1975). Quine's changing views on analyticity are surveyed in Creath (2004). The classic source for meaning externalism is Putnam (1975), which has provoked a vast literature. Some of the more significant papers are collected in Pessin and Goldberg (1996). Lau (2003) is a useful guide through the issues and literature.

Chapter 2

Dennett's distinction between the intentional, design and physical stances can be found in his 'Intentional systems' (reprinted in Dennett 1978 and in many other places). See also the papers collected in Dennett (1987). Dennett's philosophy of psychology is discussed in more detail in Chapter 6 (see the bibliography for Chapter 6 for more references).

The *locus classicus* for Marr's theory of vision is Marr (1982). The introductory chapter is reprinted in Bermúdez and Macpherson (2005). The main part of the book is challenging but rewarding. Marr's theory of vision is presented from a philosophical point of view in Chapter 4 of Sterelny (1990) and in Kitcher (1988). Peacocke (1986) argues that Marr's tripartite model of explanation needs to be supplemented with a further level. This is a difficult but rewarding paper.

The entry on Psychophysics (Algom 2002) in Nadel (2003) is a useful brief introduction. Somewhat more detailed is H. R. Schiffman's article in

Davis (2003). A far more comprehensive book-length treatment is the same author's (2001). Some of the philosophical implications of research in psychophysics are interestingly explored in Austen Clark (1993).

The modularity hypothesis was originally presented in Fodor (1983). A summary of Fodor's book, accompanied by peer commentaries, was published in *Behavioral and Brain Sciences* (BBS) in 1985 (Fodor 1985, reprinted in Fodor 1990, without the commentary). Papers exploring the empirical dimension of the modularity hypothesis are contained in Garfield (1987) and Hirschfeld and Gelman (1994). Karmiloff-Smith (1992) develops an account of modularity appropriate for developmental psychology. The idea of modularity has been much exploited by evolutionary psychologists. The "massive modularity hypothesis" is discussed further in Chapter 8 (see the references for that chapter).

The distinction between personal and subpersonal levels of explanation was originally introduced in those terms in Dennett (1969), although the basic idea goes back to Wittgenstein, if not before. Budd (1989) is a good guide to Wittgenstein's views in this area. A special issue of the journal *Philosophical Explorations*, edited by Bermúdez and Elton in 2000, offers more recent perspectives on the personal/subpersonal distinction. The term 'subdoxastic' is sometimes used for what I term the subpersonal level. Stich, 1978, proposed inferential integration and accessibility to consciousness as marks of the personal/doxastic level. J. Searle discusses accessibility to consciousness in Searle (1990a).

Eilan, McCarthy and Brewer (1993) is an exciting collection of essays on the philosophy and psychology of spatial representation. The paper by Pick explores the emergence of cognitive maps in infancy. Campbell (1994) and Bermúdez (1998, Chapter 8) discuss the role of spatial thinking in self-consciousness. Kantian themes in this area are explored in Cassam (1995).

What I call vertical explanations in section 2.3 have been extensively studied by philosophers of science, who tend to use the vocabulary of reduction (which, in my terms, is simply one type of vertical explanation). The classic model of intertheoretic reduction is in Chapter 11 of Nagel (1961). Philosophers of science have moved away from the model of strict reduction proposed by Nagel. Charles and Lennon (1992) is a useful collection of papers discussing various types of vertical explanation in a range of different areas. The ideal of reduction is bound up with views about the unity of science. Dupre (1995) takes a pessimistic view of the unity of science. Heil (2003) criticizes some of the metaphysical assumptions underlying the hierarchical conception of reality.

The views about the indispensability of personal-level commonsense psychology sketched out in section 2.3 are a distillation of arguments that are widely accepted among philosophers of mind and psychology. Classic expositions are Putnam (1960), Fodor (1975) and Pylyshyn (1981).

Chapter 3

See the bibliography for Chapter 2 for reading on vertical explanation and reduction.

The standard-bearers for the conception of the autonomous mind as I discuss it in the text are D. Dennett, D. Davidson, J. McDowell and J. Hornsby. The best guide to Dennett's views are the papers in his two collections *Brainstorms* (1978) and *The Intentional Stance* (1987), together with his more recent paper "Real patterns" (1991a). "Real patterns" is discussed in more detail in section 6.1 in the main text (see the bibliography for Chapter 6 below). Dennett's views have been extensively discussed by philosophers. The essays in Dahlbom (1993) and Brook and Ross (2002) offer useful selections. The journal *Behavioral and Brain Sciences* published a précis of *The Intentional Stance* in 1988 accompanied by commentaries from philosophers, psychologists and cognitive scientists.

Dennett's views thinking about the autonomy of commonsense psychology was influenced by his exposure to Gilbert Ryle. Ryle's *The Concept of Mind* (1949) offers an early version of the picture of the autonomous mind. Taylor (1964) explores what he takes to be the fundamental distinction between mechanistic and purposeful accounts of behavior, with particular reference to stimulus-response models of explanation. Versions of the autonomy picture can also be found in the essays collected in Haugeland (1998). Haugeland's work, like that of Dreyfus (1992), illustrates common themes between the picture of the autonomous mind and various strands in the continental tradition of philosophy.

The best source for Davidson's views are the papers collected in *Essays on Actions and Events* (1980a – new edition, 2001), particularly "Mental events" and "Psychology as philosophy". Lepore and McLaughlin (1985) contains a number of essays discussing Davidson's anomalous monism, including a very helpful essay by Kim ("Psychophysical laws", reprinted in Kim 1993). Heil and Mele (1993) is a collection focused on the metaphysics of mental causation and contains an interesting exchange between Davidson and Kim. Further references will be found in the bibliography for Chapter 6.

John McDowell's *Mind and World* (1994) is the best source for his general conception of the scope and limits of scientific understanding. Some of the issues the book raises are discussed by contributors to Smith (2002) (which also contains a response by McDowell). His understanding of the personal/subpersonal distinction is applied to one of Dennett's models of consciousness in McDowell (1994) (reprinted in McDowell 1998). McDowell discusses Davidson's anomalous monism in his 1985 (reprinted in his 1998 text). Relevant essays by Hornsby include her (1980–81) and (1986). These and other essays are reprinted in her (1997) collection.

The essays in Sosa and Tooley (1993) are a good introduction to contemporary debates about the nature of causation. Davidson (1967), reprinted in Sosa and Tooley (1993) and in his (1980a/2001), is a clear

statement of the thesis that causation requires causal laws. Schiffer (1991) argues forcefully that there are no *ceteris paribus* laws in psychology, in opposition to Lepore and Loewer (1987, 1989) and Fodor (1989). *Ceteris paribus* laws are defended in Pietrowski and Rey (1995).

The dependence of causal relations upon causal laws has been challenged, both by *singularists* such as Ducasse (see his 1926, reprinted in Sosa and Tooley 1993) and by proponents of the counterfactual theory of causation. The role of singular explanation in the social sciences is explored in Ruben (1990). The counterfactual theory of causation was first developed by Lewis (see Lewis 1973b). See also Ruben (1994). Counterfactual theorists have found it difficult to accommodate apparent counter-examples – representative discussion will be found in the essays in Collins, Hall and Paul (2004), which contains a useful extended introduction. Baker (1995) develops the counterfactual approach to mental causation.

Some prominent varieties of functionalism are carefully explained and distinguished in the first few chapters of Kim (1996) (note that what Kim calls machine functionalism falls under what I call the representational picture of the representational mind) and in Chapter 7 of Rey (1997). Lewis (1972 and 1994) are influential expositions of folk functionalism from one of its primary exponents. A more overarching version of functionalism is applied to topics in metaphysics and value theory in Jackson (2000). The *locus classicus* for *a priori* functionalism is the essays collected in Shoemaker (1984). Psychological functionalism (also known as psychofunctionalism or homuncular functionalism) is not discussed by Kim. Important book-length presentations are Lycan (1987) and Cummins (1983). See also Haugeland (1981) (published with accompanying commentaries) and Cummins (2000). Ariew, Cummins and Perlman (2002) is a recent collection of essays on functional explanation in psychology and biology.

Chapter 4

The picture of the representational mind is closely tied to the computational paradigm in artificial intelligence and cognitive science. Useful and short introductions to the computational paradigm will be found in Chapters 4 and 5 of Copeland (1993); in Chapter 2 of Dawson (1998); in the first three sections of the Introduction to Haugeland (1997); and in Chapters 1–3 of Johnson-Laird (1988). There is a careful exposition in the first two chapters of Horst (1996) (the remainder of the book is a sustained critique of the computational approach). Haugeland (1985) is an engaging introduction to artificial intelligence that expounds some of the key themes of the computational picture. Newell and Simon (1976) (reprinted in Boden 1990 and Haugeland 1997) is an influential statement of the "physical systems hypothesis". Zenon Pylyshyn's (1984) is a book-length exposition of the version of computationalism that he shares (more or less) with Fodor. Pylyshyn (1984) (reprinted in Bermúdez and Macpherson 2005) is a *BBS*

target paper published with accompanying commentaries. Fodor's own version of the representational picture is presented in numerous places. The best sources are probably *Psychosemantics* (Fodor 1987) and *The Language of Thought* (Fodor 1975). Further references to discussion of the language of thought hypothesis will be found in the bibliography for Chapters 9 and 10. Many of the philosophical motivations for the representational mind are explored in Sterelny (1990) and Crane (1995) (2nd edition, 2003). Block (1995) explores the metaphor of the mind as the software of the brain.

The computational approach to the mind is closely tied to important research in mathematical logic and the theory of computation. Rogers (1971) is an accessible introduction to the basic concepts and structures of meta-logic. Nagel and Newman (1958) gives an informal account of the significance and basic structure of Gödel's proof of the incompleteness of arithmetic. Boolos and Jeffrey (1990) is a classic introduction to the mathematical theory of computation, but readers may find Cutland (1980) easier going. Davis *et al.* (1994) presents many of the basic topics in theoretical computer science and is written for readers with a programming background.

Key presentations of the case for functional role/conceptual role semantics are Loar (1981), Block (1986), Harman (1987) (reprinted in his 1999). These authors all stress the causal dimension of functional roles. The project is critically assessed in Chapter 6 of Fodor and Lepore (1992). Field (1977) offers a version of conceptual role semantics that understands conceptual role in terms of subjective probability. Peacocke (1992) incorporates considerations of conceptual role in his theory of concepts (although he is a long way from being a functional role semanticist).

Representational theorists have a wide range of semantic theories from which to choose. Loewer (1997) is a useful chapter surveying the principal theoretical options. Cummins (1989) is a short volume that does the same job in more detail. Stich and Warfield (1994) is a useful collection of key papers in naturalized semantics. Stich himself has proposed a version of computationalism that is purely syntactic. See Stich (1983).

An important issue for computational theorists is whether computationalism is committed to a "wide" or "narrow" view of cognition. A number of theorists have argued that computationalism entails that psychological states should be individuated in terms of intrinsic physical properties of individuals, see, for example, Egan (1992) and Segal (1991, 2000). The opposite view is taken by Burge (1986, 1987) and Wilson (1994).

Chapter 5

P. S. Churchland (1986) vigorously propounds the co-evolutionary research paradigm integral to the neurocomputational picture of the mind. It was developed by Churchland and others as an alternative to standard models of intertheoretic reduction, with considerable support from studies in the

history and philosophy of science. See the bibliography for Chapter 2 for references to the standard model of intertheoretic reduction. Important works revising the standard model include Feyerabend (1962), Schaffner (1967) and Hooker (1981). The interesting case of thermodynamics and statistical mechanics is discussed in Chapter 9 of Sklar (1993) and in the same author's (1999), which is discussed in Hellman (1999). P. M. Churchland, himself a prominent neurocomputational theorist, has proposed a neurocomputational philosophy of science – see the essays collected in his 1989b. Austen Clark (1980) explores proposals to reduce psychological models to neural mechanisms. A forceful version of "psychoneural reductionism" is put forward in Bickle (1998, 2003a and 2003b). Bickle is distinctive among contemporary philosophers in thinking that molecular biology holds the key to understanding cognition. The general issue of reduction has been much discussed in contemporary philosophy of mind, although usually in abstraction from the details of how a reduction might work. J. Kim is a prominent theorist – see particularly the papers collected in Part II of his (1993).

Within philosophy the standard-bearers for the neurocomputational approach are P. M. (Paul) Churchland and P. S. (Patricia) Churchland. P. S. Churchland (1986) introduces her neurophilosophical approach to the mind and also contains much useful introductory neuroscience. The most sustained single-volume exposition of the neurocomputational approach is P. S. Churchland and T. J. Sejnowski (1992). P. M. Churchland (1995) is more philosophically motivated. The mathematically literate will learn much from Eliasmith and Andersen (2003).

The Handbook of Brain Theory and Neural Networks (Arbib 2003) is the most comprehensive single-volume source for different types of computational neuroscience and neural computing, together with entries on neuroanatomy and many other neural topics. It contains useful introductory material and "road maps". Connectionism is the form of computational neuroscience/neural computing best known to philosophers. McLeod, Plunkett and Rolls (1998) is a good introduction that comes with software allowing readers to get hands-on experience in connectionist modeling. Bechtel and Abrahamsen (1991) (2nd edition 2001) is also to be recommended. Useful article-length presentations are Rumelhart (1989) (in Posner 1989 and Haugeland 1997) and Churchland (1990) (in Cummins and Cummins 2000). Some of the potential implications of connectionism for the philosophy of mind are explored in Andy Clark (1989 and 1993). Macdonald and Macdonald (1995b) collect some key papers in this area, including the debate between Smolensky and Fodor about the structure of connectionist networks (see the bibliography for Chapter 9 for further references on this topic). Other collections include Davis (1993) and Ramsey, Stich and Rumelhart (1991).

One topic not discussed in the text is the computational power of artificial neural networks. This bears on the question of how the neurocomputa-

tional approach relates to the computational picture explored in Chapter 4. It is sometimes suggested that connectionist networks are computationally equivalent to digital computers (in virtue of being able to compute all Turing-computable functions), which might be taken to indicate that connectionist networks are simply implementations of digital computers. The implementation thesis is canvassed both by opponents of connectionism (Fodor and Pylyshyn 1988) and by leading connectionist modelers (Hinton, McClelland and Rumelhart 1986). Siegelmann and Sontag (1991) present a neural network that can simulate a universal Turing machine. Hadley (2000) expresses some skepticism about the computational assumptions involved in this and other claims about the computational power of connectionist networks.

Language acquisition has been a key area of research for neural network modeling. The pioneering study was the model of past tense acquisition proposed in Rumelhart and McClelland (1986). This model provoked considerable criticism and inspired much further research. Useful summaries will be found in Chapter 9 of McLeod, Plunkett and Rolls (1998), in Plunkett (1995) and in MacWhinney (2003). Elman *et al.* (1996) presents a connectionist perspective on development and includes a lengthy discussion of language acquisition.

The neurocomputational approach to the mind has been developed as a form of eliminative materialism (although this is by no means an integral part of the approach). The most influential presentation of eliminativism about commonsense psychology is Churchland (1981) (see also Rorty 1970 and Stich 1983). This paper has been extensively discussed. See Kitcher (1984), Horgan and Woodward (1985) and Boghossian (1990). Bermúdez forthcoming explores alternative ways of arguing for eliminativism.

Chapter 6

The classic exposition of Dennett's account of real patterns is Dennett (1991a). It is discussed in Haugeland (1993) (in Dalhbom 1993 and reprinted in Haugeland 1998), in Kirk (1993), in Nelkin (1994) and in Cohen (1995). Dennett makes much of Conway's Game of Life. There are popular expositions of the Game of Life in Poundstone (1985) and Holland (1998). Readers interested in pursuing the discussion of optimal foraging theory in the text are directed to Chapter 5 of Dawkins (1995), to Krebs and Kacelnik (1991) and to Parker and Maynard-Smith (1990) (a more technical review article in *Nature*).

A brief exposition of some of the most important experiments in the reasoning literature will be found in Chapter 1 of Goldman (1993a). Tversky and Kahneman (1974) (reprinted in Moser 1990) surveys what they consider to be cognitive biases in probabilistic and statistical reasoning resulting from reliance on heuristics. There is a book-length survey of the psychology of reasoning in Evans and Over (1996). Gerd Gigerenzer and his ABC

research group are the leading proponents of "fast and frugal heuristics". See the essays collected in Gigerenzer *et al.* (1999). A précis of the book with accompanying peer commentary was published in *BBS* (Todd and Gigerenzer 2000). The debate about whether the reasoning experiments show humans to be fundamentally irrational is surveyed in Stein (1996). Significant contributions to the debate include Cohen (1981) and Stich (1990). Proponents of the massive modularity hypothesis have also discussed the psychology of reasoning. See the references for Chapter 8.

The closest Davidson comes to giving an argument for the anomalism of the mental is in Davidson (1970, 1974). Most commentators on anomalous monism have focused on how successful Davidson is in reconciling his three principles, rather than on how plausible they are individually, e.g. Antony (1989) and the essays in Heil and Mele (1993). Exceptions include Godow (1979), Patterson (1996), Yalowitz (1997) and Tiffany (2001). The latter two discuss the argument from the uncodifiability of rationality that William Child offers on Davidson's behalf in his (1993 and 1994). Background to the discussion of decision theory and game theory in section 6.2 will be found in Allingham (2002) and Hargreaves Heap and Varoufakis (1985) (for further references to the prisoner's dilemma, see the annotated bibliography for Chapter 8).

The most influential recent version of the counterfactual theory of causation has been proposed and developed by David Lewis, particularly in his (1973b and 1979). See also Ruben (1994) for a counterfactual account of causal explanatoriness. Counterfactual theories have been extensively discussed, with particular attention to increasingly convoluted potential counterexamples. There are useful surveys of the current state of play in Menzies (2001) (in the online *Stanford Encyclopedia of Philosophy*) and, somewhat longer, in the introduction to Collins, Hall and Paul (2004). Autonomy theorists do not need to accept a counterfactual theory of causation in general. All they need is a counterfactual theory of mental causation. Ryle (1949) contains the germs of a counterfactual theory, but in the context of a behaviorism that is antithetical to standard construals of mental causation. The most worked-out theory in this area is Baker (1995). The question of how to understand counterfactuals has been extensively discussed. A very useful overview and discussion will be found in Sanford (1989). The case for possible worlds realism is made in Lewis (1986). Loux (1979) is a collection of influential essays in the metaphysics of modality that contains a number of "ersatz" accounts of possible worlds (see the essays by Stalnaker, Adams, Lycan, Cresswell, Rescher and Plantinga).

Chapter 7

The clearest statements of eliminativism about the propositional attitudes are to be found in Churchland (1981) and Stich (1983). See also Dennett (1988) for eliminativism about qualia. Stich (1996) revises his earlier elimi-

nativism. Greenwood (1991) is a very useful collection of papers on commonsense psychology, many of which deal with eliminativism. Some are published elsewhere, such as the well-known Horgan and Woodward (1985). Lynne Baker has been a vocal opponent of eliminativism. See her (1987), Chapter 7 of which is reprinted in Heil (2004).

Morton (1980) was one of the books that inspired contemporary discussion of commonsense (or folk) psychology. Morton's most recent book (Morton 2003) contains much of interest. He adopts a version of the narrow construal of commonsense psychology. There are a number of useful collections of papers on various aspects of commonsense psychology, particularly Greenwood (1991), Davies and Stone (1995a and 1995b) and Carruthers and Smith (1996). These volumes all concentrate on the debate between simulationists and theory-theorists. Gordon (1986) and Heal (1986) are key statements of the simulationist position, while more recently Currie and Ravenscroft (2002) develops a theory of imagination in the context of a simulationist approach to social understanding. Classic statements of the theory-theory approach include Fodor (1987), Churchland (1989a) and Lewis (1994). The debate between simulationists and theory-theorists has both philosophical and psychological dimensions. Carruthers and Smith (1996) includes interesting material from developmental psychologists and students of primate cognition. A number of empirical objections to the simulation theory are presented in Stich and Nichols (1992). The argument from choice effects discussed in the text is developed at length in Nichols, Stich and Leslie (1995). See further Nichols and Stich (2003).

The false belief task discussed in the text was first presented in Wimmer and Perner (1983). It turns out that children with autism are far less successful on the false belief task. A number of theorists have concluded that autism is essentially a disorder in mind-reading. For a book-length discussion of autism as "mind-blindness", see Baron-Cohen (2000). This interpretation of autism is challenged in Boucher (1996) (in Carruthers and Smith 1996). The papers in Baron-Cohen, Tager-Flusberg and Cohen (1999) discuss autism from the perspective of developmental psychology and cognitive neuroscience. The two entries on autism in Nadel (2003) provide useful background.

Miller (1997) (in Hale and Wright 1997) gives an overview of current philosophical discussions of tacit or implicit knowledge. Most discussion in this area has been inspired by claims about implicit knowledge of syntax and semantics made in linguistics. Significant contributions to the debate have been made in Evans (1981), Peacocke (1989) and Davies (1989). See Dummett (1993) for a very different approach to implicit knowledge. The modularity of commonsense psychological knowledge is discussed in Scholl and Leslie (1999) and in Segal (1996). The term "naïve physics" was first introduced in Hayes (1978) (reprinted in Boden 1990). McCloskey (1983) reviews some of the experimental work on naïve physics. A very different perspective on some of the philosophical issues raised by our naïve physics is

explored in Campbell (1994) in the context of a more wide-ranging discussion of our thinking about space, time and self-awareness. See also Peacocke (1993).

The emotions are receiving attention from philosophers, after a long period of neglect. See, for example, Griffiths (1997), Goldie (2000), Delancey (2002). The role of emotions in thinking about behavior in rational terms is discussed in Greenspan (2000). See the entry on the neural basis of emotion in Nadel (2003).

The entry on game theory in the *Stanford Encyclopaedia of Philosophy* offers a very good introduction to a complicated topic. More detailed guidance will be found in Hargreaves Heap and Varoufakis (1995). The relation between rational choice theory and commonsense psychology is discussed in Pettit (1991). Some of the philosophical issues raised by thinking about psychology through the lens of decision theory are explored in Hollis (1987 and 1996). Rational choice theory is standardly "extensional" – that is, it does not allow for a given outcome being viewed differently by different agents. This aspect of rational choice theory is criticized in two very readable books by Frederic Schick (Schick 1991 and 1997). The prisoner's dilemma is discussed in all the works mentioned, as well as in the papers collected in the useful anthology Campbell and Sowdon (1985). See also Chapter 6 of Blackburn (1998).

Chapter 8

It was for a long time a guiding assumption in empirical psychology that some combination of reflexes, innate releasing mechanisms and conditioned responses (as described in 8.1) could explain all human behavior, including such complex behaviors as language mastery. This was the view of psychological behaviorists, such as Watson and Skinner. Psychological behaviorism was overthrown during the 1950s and 1960s by what has come to be known as the cognitivist revolution in psychology. A very significant event was Chomsky's review (Chomsky 1959) of Skinner's *Verbal Behavior*, in which Chomsky forcefully argued that linguistic understanding could not possibly be explained by conditioning theory. Behaviorism has proved far more long-lasting in the study of animal behavior. The relatively new discipline of cognitive ethology (generally thought to have begun with the publication of Donald Griffin's *The Question of Animal Awareness* – Griffin 1981) attempts to understand animal behavior in the wild in terms of the concepts and categories of intentional psychology. For a sympathetic presentation of cognitive ethology from a philosophical perspective, see Dennett (1983), Allen and Bekoff (1997) and Kornblith (2002). Work in the laboratory, however, has tended to pursue a stimulus-response approach. Dickinson and Balleine (1993) and Dickinson and Shanks (1995) provide interesting discussions of the conditions under which animal behavior might require psychological explanation. Some of the methodological issues here are explored in a debate

in the journal *Mind and Language*. Heyes and Dickinson (1990) inspired a response by Allen and Bekoff (1995), to which Heyes and Dickinson (1995) is a reply.

What I have called the three-stage view of the route from perception to action is more frequently assumed than discussed. It is criticized in Hurley (1998), which is difficult but rewarding. The relation between perception and action is discussed in Andy Clark (2001). See the bibliography for Chapter 2 for readings on modularity and modular processing. There is an extensive literature on functional analysis and functional explanation in psychology and biology. See, for example, Cummins (1983) and the essays collected in Ariew, Cummins and Perlman (2002).

The "output" end has been largely neglected by philosophers, who have tended to focus on perceptual and central processing. Jeannerod (1997) is recommended as an accessible introduction to the cognitive neuroscience of action. There is a rich scientific literature on motor control. Overviews with references can be found in Miall (2003) and Jordan and Wolpert (2000). See Grush (forthcoming) for philosophical discussion.

The *MIT Encyclopedia* and the *Encyclopedia of Cognitive Science* each have a number of entries that will be useful to those wanting to find out more about the psychology and neuroscience of visual perception. Ullman (1996) is an interesting and readable book on high-level vision by one of the leaders in the field. Humphreys (1992) is a useful collection of articles on visual processing from psychology, cognitive science, neuropsychology and AI.

Mention is made in the text to the idea that perception has *nonconceptual content*. The notion of nonconceptual content is rather complex. An introduction to some of the key ideas and arguments in the debate about nonconceptual content will be found in Bermúdez (2002). A number of the central papers in this area are collected in Gunther (2003). The important concept of an affordance was developed by the perceptual psychologist J. J. Gibson. Gibson (1979) introduces some of the principal ideas of Gibson's ecological approach to visual perception.

The leading proponents of the massive modularity hypothesis are Leda Cosmides and Tooby. The essays collected in Barkow, Cosmides and Tooby (1992) offer an influential statement of some of the key ideas associated with the massive modularity hypothesis. The hypothesis has been taken up and popularized by Stephen Pinker (1997). Fodor (2000) argues against the idea that modularity might be global, as do Currie and Sterlny (2000). Peter Carruthers responds in his (2003b). Carruthers provides a philosophical defence and development of the massive modularity hypothesis in his (2003a and 2004).

An important part of the case for the massive modularity hypothesis is based on evidence from studies of human reasoning. See Cosmides and Tooby (1992) and, for a recent statement, (2000). The reasoning literature is surveyed in Evans and Over (1996). There is an extensive debate about how to interpret human rationality in the light of what is often described as a

widespread susceptibility to fallacious reasoning. Some authors have drawn the conclusion that human rationality is a myth (Stich 1990). Others have taken a more nuanced view, arguing for example that the difficulties in conditional reasoning revealed by the experimental studies reflect problems in *performance*, rather than a lack of an underlying competence (Cohen 1981).

Chapter 9

There is an extensive literature on the relation between cognitive architecture and the structure of thought. A number of important papers are collected in Macdonald and Macdonald (1995b). These include the full version of Fodor and Pylyshyn (1988) (a shortened version can be found in Haugeland 1997), together with the follow-up Fodor and McLaughlin (1990) and the article by Smolensky (Smolensky 1988) to which it is a response (also in Haugeland 1997). Smolensky (1988) was originally published in *BBS* with accompanying peer commentary. Smolensky responded to Fodor and Pylyshyn's critique in Smolensky (1991) and in a long article (Smolensky 1995) written especially for the Macdonald and Macdonald (1995). Martin Davies has produced original arguments for the language of thought hypothesis from the structure of thought. See Davies (1992 and 1998). See also Chalmers (1993), Horgan and Tienson (1992), Chapter 4 of Marcus (2001) and Cummins *et al.* (2001). Horgan and Tienson (1989 and 1996) argue that connectionist architectures approximate to classical architectures (rather than implementing them) and that because of this they are well suited to modeling cognitive processes that are governed by multiple and simultaneous "soft" constraints. Some authors have raised the question of whether classical architectures can explain systematicity in the way that Fodor and Pylyshyn assume. See Hadley (1997) and Matthews (1997).

There has been considerable work on identifying structure in artificial neural networks. In addition to the papers by Smolensky, Chalmers (1990) offers an example of a structure-sensitive network. Much of the research in this area has focused on using different statistical methods for analyzing hidden unit activation to identify similarities across different networks. See Laakso and Cottrell (2000) and Churchland (1998) for an exposition of one such similarity measure and a discussion of how it might be deployed to respond to Fodor and Pylyshyn's argument from structure. Some researchers in connectionism have developed structured connectionist models that are not fully distributed. For a brief overview and references, see Shastri (2003). Macdonald and Macdonald (1995b) also contains a number of key papers on the relation between connectionism and eliminativism. These include the influential Ramsey, Stich and Garon (1991), which argues that connectionism entails eliminativism, together with responses to that paper by Andy Clark (1989/1990) and Smolensky (1995) and a final statement of the eliminativist case by Stich and Warfield (1995). There is useful discussion of the relation between connectionism and commonsense psychology in Andy

Clark (1992), which also moots the possibility that a correct account of cognition might involve a mixed model of classical and connectionist architectures (what in the text is referred to as the two-layer response). This theme is also developed in Dennett (1991b) (see particularly Chapter 9).

Gareth Evans's statement of his Generality Constraint will be found in Evans (1982, §4.3), and Rey's generalization in Rey (1995). There is an interesting critical discussion of these types of constraint in Travis (1994). See also Millikan (1993) and Chapter 2 of Peacocke (1992).

Chapter 10

Carruthers and Boucher (1998) is a useful interdisciplinary collection of papers on the relation between thought and language. See also the entry by Schiffer in Guttenplan (1994) for a survey of possible positions on the relation between thought and language. A number of important topics in this area are discussed in Davis (2003). Although not discussed in the text, debates in the philosophy of language about the role of *communicative intentions* in the theory of meaning are very relevant to how one thinks about the relation between thought and language. One argument for the communicative conception of language is that the intentions we have in using language are what determine how it should be understood. See Jacob (1997) for an argument for the language of thought hypothesis based on a neo-Gricean theory of linguistic meaning and understanding. Avramides (1997) is a useful survey of intention-based semantics. Christopher Gauker has been a vocal critic of the communicative conception of language. See Gauker (1994 and 2003).

The inner speech hypothesis is defended in Sellars (1969), Harman (1975) and Carruthers (1996). See also Dennett (1991b). Aspects of the inner speech hypothesis can be found in Wittgenstein (see, for example, Wittgenstein (1953 §§327 ff.)). For a more detailed account of Wittgenstein's views on thought and language, see Chapters 5 and 6 of Budd (1989). Travis (2001) is a development and exploration of some of Wittgenstein's ideas about the nature of representation. Chapter 8 of Carruthers (1996) also presents a version of the rewiring hypothesis, which is defended in Mithen (1996) from an archeological perspective, in Karmiloff-Smith from a psychological perspective and in Bermúdez (2003a), Chapters 8 and 9 from a philosophical perspective. See also Dennett (1996) and Carruthers (2002), which is a *BBS* article published with a number of interesting commentaries. Annette Karmiloff-Smith has placed the idea of "representational redescription" at the core of her analysis of human development in infancy and early childhood. In her book *Beyond Modularity* she argues that the development of the human cognitive system is a process of making explicit information that is implicit in the system (Karmiloff-Smith 1992). This requires fundamental changes in cognitive architecture, the final and most important of which is generated by the emergence of language. For a brief outline of her views, see

the summary of her book that appeared as a target article in *Behavioral and Brain Sciences* (Karmiloff-Smith 1994). The accompanying commentaries provide a range of interesting critical perspectives. See also Clark and Karmiloff-Smith (1993). Andy Clark (1998) discusses how language serves as a tool to augment computational power.

The distinction between propositional and non-propositional thinking is discussed in Bermúdez (2003a). It has interesting connections with the distinction between conceptual and nonconceptual content (see the essays in Gunther 2003), as well as with debates about the nature of visual imagery. Many of the experiments in the imagery debate are reported in Shepard and Cooper (1982). For a shorter introduction, see the entry on Imagery in Nadel (2003) (Wraga and Kosslyn 2003). Block (1981) and Tye (1991) are book-length treatments of the so-called "imagery debate". Kosslyn (1994) offers the perspective of a leading cognitive psychologist.

The evaluation of arguments for the language of thought hypothesis depends upon how one views the cognitive abilities of non-linguistic creatures. The references in the bibliography for Chapter 8 to literature on cognitive ethology are relevant here. See Bermúdez (2003a) for a book-length treatment of the nature and limits of non-linguistic thought. Davidson (1975) (discussed in Chapter 6 of Heil 1992) is a well-known argument that non-linguistic creatures cannot have beliefs. The essays in Weiskrantz (1988) cover much of the relevant empirical material.

The argument that practical reasoning presupposes a language of thought can be found in Maloney (1989) and Rey (1997), in addition to Fodor (1975). The references on optimal foraging theory in the bibliography for Chapter 6 are relevant here. Gibson (1979) is the key presentation of his ideas about visual perception. These ideas are fiercely criticized in Fodor and Pylyshyn (1981). Dummett's conception of proto-thoughts is developed in his (1993), discussed in Chapter 3 of Bermúdez (2003a). Related ideas can be found in John Campbell's discussion of what he terms "causally indexical understanding". See Campbell (1994). Hurley and Nudds (2005) is an interdisciplinary collection of essays focused on animal decision-making and "rationality".

The approach to perception presupposed by Fodor's argument from perceptual integration has been adopted by an influential school in the psychology of perception. Influential exponents of the "unconscious inference" approach to perception include Richard Gregory and Irving Rock. See Gregory (1997) and Rock (1983) for book-length treatments. A much briefer introduction will be found in the overview article on perception in Nadel (2003) (Pomerantz 2003). Shepard (1994/2001) succinctly summarizes the case that principles reflecting the basic structure of objects are hard-wired into the brain. The papers accompanying the (2001) reprint provide further discussion of this hypothesis. There is a fascinating account of mammalian vision in Chapter 4 of Churchland and Sejnowski (1992). The problems posed by stereo vision, including the correspondence problem are explored on pp. 188–221.

There is an extensive philosophical literature on the nature of concepts and their relation to linguistic meaning. The most systematic work in this area has been done by Christopher Peacocke. See his (1992), which presents a view of concepts very different from that motivating Fodor and other language of thought theorists. Many of the psychological theories of concepts are surveyed and discussed from a philosophical point of view in Prinz (2002) and from a psychological perspective in Murphy (2002). Margolis and Laurence (1999) is a useful interdisciplinary anthology.

Harman (1970) discusses the argument that language learning requires a language of thought. Field (1978) proposes a hybrid position, according to which natural language propositional attitudes get grafted onto a more primitive language of thought available to prelinguistic children. The expression "full-blooded" as applied to theories of meaning comes from Michael Dummett, who famously criticized Davidsonian theories of meaning for failing to provide theories of understanding (Dummett 1975). Dummett's arguments provoked a response from John McDowell (McDowell 1987, reprinted in McDowell 1998).

The most comprehensive statement of Dummett's philosophical outlook is Dummett (1973), but the essays in Dummett (1981) provide a more accessible introduction, as does his well-known essay "What is a theory of meaning? (II)" (Dummett 1976), originally published in Evans and McDowell (1976/1982). An introduction to some key themes in Dummett's thinking about language will be found in Weiss (2002). A difficult argument to the effect that truth-rules of the type envisaged by Fodor cannot explain what it is to understand a language will be found in the introduction to Evans and McDowell (1976).

There has been considerable research on connectionist models of language-learning. The pioneering study was Rumelhart and McClelland (1986), reporting data presented in Brown (1973) and Kuczaj (1977). Much subsequent work in connectionist modeling was responding to critiques of this approach (such as those in Pinker and Prince (1988) and Prince and Pinker (1988). For reviews, see Plunkett (1995) and Chapter 9 of McLeod, Plunkett and Rolls (1998). Elman *et al.* (1996) is an influential argument for a connectionist perspective on development, which contains much discussion of language acquisition, particularly in Chapter 3.

Bibliography

Algom, D. (2002) "Psychophysics", in L. Nadel (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.

Allen, C. and Bekoff, M. (1995) "Cognitive Ethology and the Intentionality of Animal Behaviour", *Mind and Language*, 10(4): 313–328.

—— (1997) *Species of Mind*, Cambridge, MA: MIT Press.

Allingham, M. (2002) *Choice Theory: A Very Short Introduction*, Oxford: Oxford University Press.

Antony, L. (1989) "Anomalous Monism and the Problem of Explanatory Force", *Philosophical Review*, 98: 153–187.

Arbib, M. (ed.) (2003) *The Handbook of Brain Theory and Neural Networks*, 2nd edn, Cambridge, MA: MIT Press.

Ariew, A., Cummins, R. and Perlman, M. (eds) (2002) *Functions: New Essays in the Philosophy of Psychology and Biology*, Oxford: Oxford University Press.

Armstrong, D. M. (1962) *Bodily Sensations*, London: Routledge and Kegan Paul.

Avramides, A. (1989) *Meaning and Mind: An Examination of the Gricean Account of Language*, Cambridge, MA: MIT Press.

Axelrod, R. (1984) *The Evolution of Cooperation*, Harmondsworth: Penguin.

Baddeley, A. D. (1998) *Human Memory: Theory and Practice*, revised edn, Boston: Allyn and Bacon.

Baddeley, A. D. and Hitch, G. J. (1974) "Working Memory", in G. Bower (ed.) *The Psychology of Learning and Motivation*, vol. VIII, New York: Academic Press, pp. 47–89.

Baker, L. R. (1987) *Saving Belief: A Critique of Physicalism*, Princeton, NJ: Princeton University Press.

—— (1995) *Explaining Attitudes: A Practical Approach to the Mind*, Cambridge: Cambridge University Press.

Barkow, J. H., Cosmides, L. and Tooby, J. (eds) (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, New York: Oxford University Press.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*, Cambridge, MA: MIT Press.

Baron-Cohen, S., Tager-Flusberg, H. and Cohen, D. J. (2000) *Understanding Other Minds: Perspectives from Autism*, Oxford: Oxford University Press.

Barrett, M. (1995) "Early Lexical Development", in P. Fletcher and P. MacWhinney (eds) *The Handbook of Child Language*, Malden, MA: Basil Blackwell.

Bechtel, W. and Abrahamsen, A. (1991) *Connectionism and the Mind*, Oxford: Blackwell, 2nd edn 2001.

Bechtel, W. and Graham, G. (eds) (1998) *A Companion to Cognitive Science*, Malden, MA: Blackwell.

Bechtel, W., Mandik, P., Mundale, J. and Stufflebeam, R. S. (eds) (2001) *Philosophy and the Neurosciences: A Reader*, Malden, MA: Blackwell.

Berkeley, G. (1975) *Philosophical Works*, ed. M. R. Ayers, London: Dent.

Bermúdez, J. L. (1998) *The Paradox of Self-Consciousness*, Cambridge, MA: MIT Press.
—— (2002) "Nonconceptual Content", in L. Nadel (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.
—— (2003a) *Thinking Without Words*, New York: Oxford University Press.
—— (2003b) "Nonconceptual Mental Content", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Spring edition)*, url = <http://plato.stanford.edu/archives/spr2003/entries/content-nonconceptual/>.
—— (forthcoming) "Arguing for Eliminativism", in B. Keeley (ed.) *Paul Churchland*, Cambridge: Cambridge University Press.
Bermúdez, J. L. and Elton, M. E. (eds) (2000) *Personal and Subpersonal: Essays on Psychological Explanation, Philosophical Explorations*, 3(1): pp. 1–119.
Bermúdez, J. L. and Macpherson, F. (eds) (2005) *Philosophy of Psychology: Contemporary Readings*, London: Routledge.
Bickle, J. (1998) *Psychoneural Reduction: The New Wave*, Cambridge, MA: MIT Press.
—— (2003a) *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Dordrecht: Kluwer Academic Publishers.
—— (2003b) "Philosophy of Mind and the Neurosciences", in S. Stich and T. Warfield (eds) *Blackwell Guide to Philosophy of Mind*, New York: Basil Blackwell.
Blackburn, S. (1998) *Ruling Passions*, Oxford: Oxford University Press.
Block, N. (ed.) (1980) *Readings in Philosophy of Psychology*, Cambridge, MA: Harvard University Press.
—— (1981) *Imagery*, Cambridge, MA: MIT Press.
—— (1986) "Advertisement for a Semantics for Psychology", in P. French, T. Uehling and H. Wettstein (eds) *Midwest Studies in Philosophy*, Minneapolis: University of Minnesota Press, vol. X: pp. 615–678.
—— (1995) "The Mind as the Software of the Brain", in D. Osherson, L. Gleitman, S. Kosslyn, E. Smith and S. Sternberg (eds) *An Invitation to Cognitive Science,* 2nd edn, Cambridge, MA: MIT Press.
Block, N. and Stalnaker, R. (1999) "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review*, 108: 1–46.
Boden, M. (ed.) (1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.
Boghossian, P. A. (1990) "The Status of Content", *Philosophical Review*, 99(2): 157–184.
—— (1997) "Analyticity", in B. Hale and C. Wright (eds) *The Philosophy of Language*, Oxford: Basil Blackwell.
Boolos, G. and Jeffrey, R. (1990) *Computability and Logic*, 3rd edn, Cambridge: Cambridge University Press.
Botterill, G. and Carruthers, P. (1999) *The Philosophy of Psychology*, Cambridge: Cambridge University Press.
Boucher, J. (1996) "What Could Possibly Explain Autism?", in P. Carruthers and P. K. Smith (eds) *Language and Thought*, Cambridge: Cambridge University Press.
Braddon-Mitchell, D. and Jackson, F. (1996) *Philosophy of Mind and Cognition*, Oxford: Blackwell.
Brandom, R. (2000) *Articulating Reasons: An Introduction to Inferentialism*, Cambridge, MA: Harvard University Press.
Bressler, S. L. (2003) "Event-related Potentials", in M. A. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press.

Brook, A. (1994) *Kant and the Mind*, New York: Cambridge University Press.
Brown, R. (1973) *A First Language: The Early Stages*, Cambridge, MA: Harvard University Press.
Bruce, V. and Young, A. (1986) "Understanding Face Recognition," *British Journal of Psychology*, 77: 305–327.
Buckner, R. L. and Petersen, S. E. (1998) "Neuroimaging", in W. Bechtel and G. Graham (eds).
Budd, M. (1989) *Wittgenstein's Philosophy of Psychology*, London: Routledge.
Burge, T. (1982) "Other Bodies", in A. Woodfield (ed.) *Thought and Object: Essays on Intentionality*, Oxford: Clarendon Press.
—— (1986) "Individualism and Psychology", *Philosophical Review*, 95(1): 3–45.
—— (1987) "Marr's Theory of Vision", in L. Garfield (ed.) *Modularity in Knowledge Representation and Natural Language Understanding*, Cambridge, MA: MIT Press.
Callender, C. (1999) "Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy", *Journal of Philosophy*, 96(7): 348–373.
Campbell, J. (1994) *Past, Space and Self*, Cambridge, MA: MIT Press.
Campbell, R. and Sowdon, L. (1985) *Paradoxes of Cooperation and Rationality*, Vancouver, BC: University of British Columbia Press.
Caramazza, A. (1986) "On Drawing Inferences about the Structure of Normal Cognitive Systems from the Analysis of Patterns of Impaired Performance: The Case for Single-Patient Studies", *Brain and Cognition*, 5: 41–66.
Carruthers, P. (1996) *Language, Thought and Consciousness*, Cambridge: Cambridge University Press.
—— (2002) "Modularity, Language, and the Flexibility of Thought", *Behavioral and Brain Sciences*, 25(6): 657–674.
—— (2003a) "Moderately Massive Modularity", in A. O'Hear (ed.) *Mind and Persons*, Cambridge: Cambridge University Press.
—— (2003b) "On Fodor's Problem", *Mind and Language*, 18(5): 502–523.
—— (2004) "Practical Reason in a Modular Mind", *Mind and Language*, 19(3): 259–278.
Carruthers, P. and Boucher, J. (eds) (1998) *Language and Thought: Interdisciplinary Themes*, Cambridge: Cambridge University Press.
Carruthers, P. and Smith, P. K. (eds) (1996) *Theories of Theories of Mind*, Cambridge: Cambridge University Press.
Cartwright, N. (1983) *How the Laws of Physics Lie*, Oxford: Clarendon Press.
Cassam, Q. (1995) "Introspection and Bodily Self-Ascription", in J. L. Bermúdez, A. Marcel and N. Eilan (eds) *The Body and the Self*, Cambridge, MA: MIT Press.
Chalmers, D. (1990) "Syntactic Transformations on Distributed Representations", *Connection Science*, 2: 53–62.
—— (1993) "Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong", *Philosophical Psychology*, 6: 305–319.
—— (1996) *The Conscious Mind*, New York: Oxford University Press.
—— (ed.) (2002) *Philosophy of Mind: Classical and Contemporary Readings*, New York: Oxford University Press.
Chalmers, D. and Jackson, F. (2001) "Conceptual Analysis and Reductive Explanation", *Philosophical Review*, 110: 315–360.
Changeux, J. P. (1985) *Neuronal Man: The Biology of Mind*, Oxford: Oxford University Press.

Charles, D. and Lennon, K. (eds) (1992) *Reduction, Explanation, Realism*, Oxford: Clarendon Press.

Child, W. (1993) "Anomalism, Uncodifiability, and Psychophysical Relations", *Philosophical Review*, 102(2): 215–245.

—— (1994) *Causality, Interpretation, and the Mind*, Oxford: Clarendon Press.

Chomsky, N. (1959) "A Review of B. F. Skinner's *Verbal Behavior*", *Language*, 35(1): 26–58.

—— (1980) *Rules and Representations*, New York: Columbia University Press.

Churchland, P. M. (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press.

—— (1981) "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy*, 78–2: 67–90, reprinted in W. Lycan (ed.) (1990) *Mind and Cognition: A Reader*, Oxford: Basil Blackwell.

—— (1988) *Matter and Consciousness*, revised edn, Cambridge, MA: MIT Press.

—— (1989a) "Folk Psychology and the Explanation of Human Behavior", *Philosophical Perspectives*, 3: 225–241.

—— (1989b) *A Neurocomputational Perspective*, Cambridge, MA: MIT Press.

—— (1990) "Cognitive Activity in Artificial Neural Networks", in D. Osherson and E. Smith (eds) *Thinking*, Cambridge, MA: MIT Press, reprinted in D. Cummins and R. Cummins (eds) *Minds, Brains and Computers*, Oxford: Blackwell, 2000.

—— (1992) "A Deeper Unity: Some Feyerabendian Themes in Neurocomputational Form", in S. Davis (ed.) *Connectionism: Theory and Practice*, New York: Oxford University Press.

—— (1995) *The Engine of Reason, the Seat of the Soul*, Cambridge, MA: MIT Press.

—— (1998) "Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered", *Journal of Philosophy*, 95: 5–32.

Churchland, P. S. (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge, MA: MIT Press.

Churchland, P. S. and Sejnowski, T. J. (1992) *The Computational Brain*, Cambridge, MA: MIT Press.

Clark, A. and Karmiloff-Smith, A. (1993) "The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought", *Mind and Language*, 8–3: 488–519.

Clark, Austen (1980) *Psychological Models and Neural Mechanisms: An Examination of Reductionism in Psychology*, Oxford: Oxford University Press.

—— (1993) *Sensory Qualities*, Oxford: Clarendon Press.

Clarke, Andy (1989) *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, MA: MIT Press.

—— (1989/1990) "Connectionist Minds", *Proceedings of the Aristotelian Society*, 90: 83–102.

—— (1993) *Associative Engines: Connectionism, Concepts and Representational Change*, Cambridge, MA: MIT Press.

—— (1997) *Being There: Putting Brain, Body and World Together Again*, Cambridge, MA: MIT Press.

—— (1998) "Magic Words: How Language Augments Human Condition", in P. Carruthers and J. Boucher (eds) *Language and Thought: Interdisciplinary Themes*, Cambridge: Cambridge University Press.

—— (2001a) *Mindware: An Introduction to the Philosophy of Cognitive Science*, Oxford: Oxford University Press.

—— (2001b) "Visual Experience and Motor Action: Are the Bonds Too Tight?", *Philosophical Review*, 110(4): 495–519.
—— (2003) *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*, Oxford: Oxford University Press.
Clarke, D. (2003) *Descartes's Theory of Mind*, Oxford: Oxford University Press.
Cohen, B. (1995) "Patterns Lost: Indeterminacy and Dennett's Realism about Beliefs", *Pacific Philosophical Quarterly*, 76(1): 17–31.
Cohen, J. (1981) "Can Human Irrationality Be Experimentally Demonstrated?", *Behavioral and Brain Sciences*, 4: 317–370.
Collins, J., Hall, E. and Paul, L. (2004) *Causation and Counterfactuals*, Cambridge, MA: MIT Press.
Copeland, J. (1993) *Artificial Intelligence: A Philosophical Introduction*, Oxford: Blackwell.
Cosmides, L. (1989) "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task", *Cognition*, 31: 187–276.
Cosmides, L. and Tooby, J. (1992) "Cognitive Adaptations for Social Exchange", in J. Barkow, L. Cosmides and J. Tooby (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, New York: Oxford University Press.
—— (1994) "Origins of Domain Specificity: The Evolution of Functional Organization", in L. Hirschfeld and S. Gelman (eds) *Mapping the Mind: Domain Specificity in Cognition and Culture*, New York: Cambridge University Press.
—— (2000) "The Cognitive Neuroscience of Social Reasoning", in M. Gazzaniga (ed.) *The New Cognitive Neurosciences*, Cambridge, MA: MIT Press.
Cowie, R. (1977) "Optimal Foraging in Great Tits (Parus Major)", *Nature*, 268: 137–139.
Crane, T. (1995) *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*, 2nd edn, Harmondsworth: Penguin, 2003.
Creath, R. (2004) "Quine on the Intelligibility and Relevance of the Analytic", in R. F. Gibson (ed.) *The Cambridge Companion to Quine*, Cambridge: Cambridge University Press.
Cummins, R. (1983) *The Nature of Psychological Explanation*, Cambridge, MA: MIT Press.
—— (1989) *Meaning and Mental Representation*, Cambridge, MA: MIT Press.
—— (2000) "How Does it Work? vs. What are the Laws? Two Conceptions of Psychological Explanation", in F. Keil and R. Wilson (eds) *Explanation and Cognition*, Cambridge, MA: MIT Press.
Cummins, R. and Cummins, D. (eds) (1999) *Minds, Brains, and Computers*, New York: Oxford University Press.
Cummins, R. *et al.* (2001) "Systematicity and the Cognition of Structured Domains", *Journal of Philosophy*, 98: 167–185.
Currie, G. (1995) "Imagination and Simulation", in M. K. Davies and T. Stone (eds) *Mental Simulation*, Oxford: Blackwell.
Currie, G. and Ravenscroft, I. (2002) *Recreative Minds: Thought, Imagination and Perception*, Oxford: Oxford University Press.
Currie, G. and Sterelny, K. (2000) "How to Think about the Modularity of Mind Reading", *Philosophical Quarterly*, 50(199): 145–160.
Cutland, N. (1980) *Computability and Logic*, Cambridge: Cambridge University Press.

Darwin, C. (1872) *The Expression of the Emotions in Man and Animals*, London: John Murray.
Davidson, D. (1967) "Truth and Meaning", *Synthese*, 17: 304–323, reprinted in Davidson (1980b).
—— (1969) "Actions, Reasons and Causes", *Journal of Philosophy*, 60: 685–700, reprinted in Davidson (1980a).
—— (1970) "Mental Events", in L. Foster and J. Swanson (eds) *Experience and Theory*, Amherst, MA: University of Massachusetts Press, pp. 207–227. References are to the reprinted version in Davidson (1980a).
—— (1971) "Agency", in R. Binkley, R. Bronaugh and A. Marras (eds) *Agent, Action, and Reason*, Toronto: University of Toronto Press. References are to the reprinted version in Davidson (1980a).
—— (1974) "Psychology as Philosophy", in S. Brown (ed.) *Philosophy of Psychology*, London: Macmillan, pp. 41–52. References are to the reprinted version in Davidson (1980a).
—— (1975) "Talk and Thought", in S. Guttenplan (ed.) *Mind and Language*, Oxford: Oxford University Press.
—— (1980a) *Essays on Actions and Events*, Oxford: Clarendon Press.
—— (1980b) *Essays on Truth and Interpretation*, Oxford: Clarendon Press.
Davies, M. (1986) "Tacit Knowledge, and the Structure of Thought and Language", in C. Travis (ed.) *Meaning and Interpretation*, Oxford: Blackwell.
—— (1989) "Tacit Knowledge and Subdoxastic States", in A. George (ed.) *Reflections on Chomsky*, Oxford: Blackwell.
—— (1992) "Aunty's Own Argument for the Language of Thought", in J. Ezquerro and J. Larrazabal (eds) *Cognition, Semantics and Ontology*, Norwell, MA: Kluwer.
—— (1998) "Language, Thought, and the Language of Thought: Aunty's Own Argument Revisited", in P. Carruthers and J. Boucher (eds) *Language and Thought: Interdisciplinary Themes*, Cambridge: Cambridge University Press.
Davies, M. K. and Stone, T. (eds) (1995a) *Folk Psychology: The Theory of Mind Debate*, Oxford: Blackwell.
—— (eds) (1995b) *Mental Simulation*, Oxford: Blackwell.
Davis, M. D. *et al.* (1994) *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, New York: Academic Press.
Davis, S. (1993) *Connectionism: Theory and Practice*, New York: Oxford University Press.
Davis, S. F. (ed.) (2003) *Handbook of Research Methods in Experimental Psychology*, Malden, MA: Blackwell.
Davis, W. A. (2003) *Meaning, Expression, and Thought*, Cambridge: Cambridge University Press.
Dawkins, M. S. (1995) *Unravelling Animal Behaviour*, 2nd edn, London: Longman.
Dawkins, R. (1989) *The Selfish Gene*, new edn, Oxford: Oxford University Press.
Dawson, M. R. W. (1998) *Understanding Cognitive Science*, Oxford: Blackwell.
DeLancey, C. (2002) *Passionate Engines: What Emotions Reveal about Mind and Artificial Intelligence*, New York: Oxford University Press.
Dennett, D. (1969) *Content and Consciousness*, London: Routledge.
—— (1971) "Intentional Systems", *Journal of Philosophy*, 68: 87–106, reprinted in Dennett (1978).
—— (1975) "Brain Writing and Mind Reading", in K. Gunderson (ed.) *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science*, vol. VII,

University of Minnesota Press. Reprinted, with postscript in D. Rosenthal (ed.) *The Nature of Mind*, Oxford: Oxford University Press, 1991.
—— (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery, VT: Bradford Books, reprinted in 1981.
—— (1981) "True Believers: The Intentional Strategy and Why it Works", in A. F. Heath (ed.) *Scientific Explanation*, Oxford: Oxford University Press, reprinted in W. Lycan (ed.) *Mind and Cognition*, Oxford: Blackwell, 1990.
—— (1983) "Intentional Systems in Cognitive Ethology: the 'Panglossian Paradigm' Defended", *Behavioral and Brain Sciences*, 6: 343–390.
—— (1984) "Cognitive Wheels: The Frame Problem of AI", in C. Hookway (ed.) *Minds, Machines and Evolution*, Cambridge: Cambridge University Press.
—— (1987) *The Intentional Stance*, Cambridge: Cambridge University Press.
—— (1988) "Quining Qualia", in A. Marcel and E. Bisiach (eds) *Consciousness in Modern Science*, Oxford: Oxford University Press, reprinted in W. Lycan (ed.) *Mind and Cognition*, Oxford: Blackwell, 1990.
—— (1991a) "Real Patterns", *Journal of Philosophy*, 88: 27–51, reprinted in W. Lycan (ed.) *Mind and Cognition*, Oxford: Blackwell, 1990, 2nd edn.
—— (1991b) *Consciousness Explained*, Boston: Little, Brown & Co.
—— (1995) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, New York: Simon and Schuster.
—— (1996) *Kinds of Minds*, New York: Basic Books.
Dickinson, A. (1980) *Contemporary Animal Learning Theory*, Cambridge: Cambridge University Press.
Dickinson, A. and Balleine, B. (1993) "Actions and Responses: The Dual Psychology of Behavior", in N. Eilan, B. Brewer and R. McCarthy (eds) *Spatial Representations*, Oxford: Blackwell.
Dickinson, A. and Shanks, D. (1995) "Instrumental Action and Causal Representation", in D. Sperber, D. Premack and A. J. Premack (eds) *Causal Cognition*, New York: Oxford University Press.
Donald, M. (1991) *Origins of the Modern Mind*, Cambridge, MA: Harvard University Press.
Dreyfus, H. L. (1992) *What Computers Still Can't Do*, 3rd edn, Cambridge, MA: MIT Press.
Ducasse, C. J. (1926) "On the Nature and Observability of the Causal Relation", *Journal of Philosophy*, 23: 57–68, reprinted in E. Sosa and M. Tooley (eds) *Causation*, Oxford: Oxford University Press, 1993.
Dummett, M. (1973) *Frege: Philosophy of Language*, London: Duckworth.
—— (1975) "What is a Theory of Meaning? I", in S. Guttenplan (ed.) *Mind and Language*, Oxford: Oxford University Press.
—— (1976) "What is a Theory of Meaning? II", in G. Evans and J. McDowell (eds) *Truth and Meaning*, Oxford: Clarendon Press.
—— (1981) *The Interpretation of Frege's Philosophy*, London: Duckworth, and Cambridge, MA: Harvard University Press.
—— (1993) "What Do I Know When I Know a Language?", in *The Seas of Language*, Oxford: Oxford University Press.
Dupre, J. (1995) *Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Cambridge, MA: Harvard University Press.
Ebbesson, S. O. E. (1984) "Evolution and Ontogeny of Neural Circuits", *Behavioral and Brain Sciences,* 7: 321–331.

Edelman, G. (1989) *The Remembered Present: A Biological Theory of Consciousness*, New York: Basic Books.
Egan, F. (1992) "Individualism, Computation, and Perceptual Content", *Mind*, 101: 443–459.
Eilan, N., McCarthy, R. and Brewer, B. (eds) (1993) *Spatial Representation*, Oxford: Blackwell.
Eliasmith, C. and Anderson, C. H. (2003) *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*, Cambridge, MA: MIT Press.
Elman, J. L. *et al.* (1996) *Rethinking Innateness: A Connectionist Perspective on Development*, Cambridge, MA: MIT Press.
Evans, G. (1981) "Semantic Theory and Tacit Knowledge", in J. McDowell (ed.) *Gareth Evans: Collected Papers*, Oxford: Clarendon Press.
—— (1982) *The Varieties of Reference*, Oxford: Oxford University Press.
Evans, G. and McDowell, J. (eds) (1976) *Truth and Meaning*, Oxford: Clarendon Press.
Evans, J. and Over, D. E. (1996) *Rationality and Reasoning*, East Sussex: Psychology Press.
Farah, M. J. (1994) "Neuropsychological Inference with an Interactive Brain: A Critique of the 'Locality Assumption' ", *Behavioral and Brain Sciences*, 17: 43–61.
Farah, M. J. and McClelland, J. L. (1991) "A Computational Model of Semantic Memory Impairment: Modality-Specificity and Emergent Category-Specificity", *Journal of Experimental Psychology (General)*, 120(4): 339–357.
Feldmann, J. A. and Ballard, D. H. (1982) "Connectionist Models and their Properties", *Cognitive Science*, 6: 205–254.
Feyerabend, P. R. (1962) "Explanation, Reduction, and Empiricism", in H. Feigl and C. Maxwell (eds) *Minnesota Studies in the Philosophy of Science*, vol. III, University of Minnesota Press, Minneapolis.
Field, H. (1977) "Logic, Meaning and Conceptual Role", *Journal of Philosophy*, 69: 379–408.
—— (1978) "Mental Representation", *Erkenntnis* 13(1): 9–61.
Finger, S. (1994) *The Origins of Neuroscience: A History of Explorations into Brain Function*, New York: Oxford University Press.
—— (2000) *Minds Behind the Brain: A History of the Pioneers and their Discoveries*, New York: Oxford University Press.
Flanagan, O. (1984) *The Science of the Mind*, Cambridge, MA: MIT Press.
Flew, A. (ed.) (1956) *Essays in Conceptual Analysis*, London: Macmillan.
Fodor, J. A. (1975) *The Language of Thought*, New York: Crowell.
—— (1983) *The Modularity of Mind*, Cambridge, MA: MIT Press.
—— (1985) "Precise of 'Modularity of Mind' ", *Behavioral and Brain Sciences*, 8: 1–42.
—— (1986) "Why Paramecia Don't Have Mental Representations", in P. French, T. Uehling, Jr. and H. Wettstein (eds) *Midwest Studies in Philosophy*, Minneapolis: University of Minnesota Press.
—— (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA: MIT Press.
—— (1989) "Making Mind Matter More", *Philosophical Topics*, 18: 59–79.
—— (1990) *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.
—— (2000) *The Mind Doesn't Work That Way*, Cambridge, MA: MIT Press.
Fodor, J. A. and Lepore, E. (1992) *Holism: A Shopper's Guide*, Oxford: Basil Blackwell.

Fodor, J. A. and McLaughlin, B. (1990) "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work", *Cognition*, 35: 183–204.
Fodor, J. A. and Pylyshyn, Z. (1981) "How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological' Approach", *Cognition*, 9: 139–196.
—— (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, 28: 3–71. References are to the reprinted version in C. Macdonald and G. Macdonald (eds) *Connectionism: Debates on Psychological Explanation*, vol. 2, Oxford: Basil Blackwell, 1995.
Gallistel, C. R. (1990) *The Organization of Learning*, Cambridge, MA: MIT Press.
Gardner, H. (1985) *The Mind's New Science: A History of the Cognitive Revolution*, New York: Basic Books.
—— (1993) *Multiple Intelligences: The Theory in Practice*, New York: Basic Books.
—— (2000) *The Disciplined Mind: Beyond Facts and Standardized Tests, the K-12 Education That Every Child Deserves*, New York: Penguin Putnam.
Garfield, J. L. (ed.) (1987) *Modularity in Knowledge Representation and Natural-Language Understanding*, Cambridge, MA: MIT Press.
Garrett, M. F. (2003) "Language and Brain", in L. Nadel (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.
Garson, G. D. (1998) *Neural Network Analysis for Social Scientists*, London: Sage Publications.
Gauker, C. (1994) *Thinking Out Loud: An Essay on the Relation Between Thought and Language*, Princeton, NJ: Princeton University Press.
—— (2003) *Words Without Meaning*, Cambridge, MA: MIT Press.
Gauthier, D. (1986) *Morals by Agreement*, Oxford: Oxford University Press.
Gazzaniga, M. S. (ed.) (2000) *The New Cognitive Neurosciences*, 2nd edn, Cambridge, MA: MIT Press.
Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*, Boston: Houghton Mifflin.
Gigerenzer, G. *et al.* (1999) *Simple Heuristics That Make Us Smart*, New York: Oxford University Press.
Godow, A. Jr. (1979) "Davidson and the Anomalism of the Mental", *Southern Journal of Philosophy*, 17(2): 163–174.
Goldie, P. (2000) *The Emotions*, Oxford: Oxford University Press.
Goldman, A. (1976) "Discrimination and Perceptual Knowledge", *Journal of Philosophy*, 73: 771–791.
—— (1993a) *Philosophical Applications of Cognitive Science*, Boulder, CO: Westview Press.
—— (ed.) (1993b) *Readings in Philosophy and Cognitive Science*, Cambridge, MA: MIT Press.
Gopnik, A. and Meltzoff, A. (1997) *Thoughts, Theories and Things*, Cambridge, MA: MIT Press.
Gordon, R. (1986) "Folk Psychology as Simulation", *Mind and Language*, 1: 158–171, reprinted in M. K. Davies and T. Stone (eds) *Folk Psychology*, Oxford: Blackwell, 1995.
Gorman, R. P. and Sejnowski, T. J. (1988) "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets", *Neural Networks*, 1: 75–89.
Greenspan, P. (2000) "Emotional Strategies and Rationality", *Ethics*, 110: 469–487.
Greenwood, J. (ed.) (1991) *The Future of Folk Psychology*, Cambridge: Cambridge University Press.

Gregory, R. L. (1997) *Mirrors in Mind*, Harmondsworth: Penguin.
Grice, H. P. and Strawson, P. F. (1956) "In Defense of a Dogma", *Philosophical Review*, 65: 141–158.
Griffin, D. (1981) *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*, New York: Rockefeller University Press.
Griffiths, P. E. (1997) *What Emotions Really Are*, Chicago: Chicago University Press.
Griggs, R. A. and Cox, J. R. (1982) "The Elusive Thematic-materials Effect in Wason's Selection Task", *British Journal of Psychology*, 73: 407–420.
Grush, Rick (forthcoming) "The Emulation Theory of Representation: Motor Control, Imagery, and Perception", *Behavioral and Brain Sciences*.
Gunther, Y. (ed.) (2003) *Essays on Nonconceptual Content*, Cambridge, MA: MIT Press.
Guttenplan, S. (ed.) (1994) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.
Hadley, R. F. (1997) "Cognition, Systematicity and Nomic Necessity", *Mind and Language*, 12(2): 137–153.
—— (2000) "Cognition and the Computational Power of Connectionist Networks", *Connection Science*, 12(2): 95–110.
Hargreaves Heap, S. P. and Varoufakis, Y. (1995) *Game Theory: A Critical Introduction*, London: Routledge and Kegan Paul.
Harman, G. (1970) "Language learning", *Nous*, 4: 33–43. Revised version in Harman (1999).
—— (1972) "Logical Form", *Foundations of Language*, 9: 38–65.
—— (1973) *Thought*, Princeton, NJ: Princeton University Press.
—— (1975) "Language, Thought, and Communication", in K. Gunderson (ed.) *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science*, vol. VII, Minneapolis: University of Minnesota Press, reprinted in Harman (1999).
—— (1987) "(Nonsolipsistic) Conceptual Role Semantics", in E. Lepore (ed.) *Semantics of Natural Language*, New York: Academic Press.
—— (1990) "The Intrinsic Quality of Experience", *Philosophical Perspectives*, 4: 31–52.
—— (1999) *Reasoning, Meaning and Mind*, Oxford: Clarendon Press.
Harnish, R. M. (2002) *Minds, Brains, and Computers: A Historical Introduction to the Foundations of Cognitive Science*, Oxford: Blackwell.
Harnish, R. M. *et al.* (2001) *Linguistics: An Introduction to Language and Communication*, 5th edn, Cambridge, MA: MIT Press.
Hatfield, G. (1990) *The Natural and the Normative: Theories of Spatial Perception from Kant to Helmholtz*, Cambridge, MA: MIT Press.
Haugeland, J. (ed.) (1981) *Mind Design*, Cambridge, MA: MIT Press.
—— (1985) *Artificial Intelligence: The Very Idea*, Cambridge, MA: MIT Press.
—— (1993) "Pattern and Being", in B. Dahlbom (ed.) *Dennett and His Critics*, Oxford: Blackwell.
—— (ed.) (1997) *Mind Design II*, Cambridge, MA: MIT Press.
—— (1998) *Having Thought: Essays in the Metaphysics of Mind*, Cambridge, MA: Harvard University Press.
Hayes, P. (1978) "The Naive Physics Manifest", in D. Michie (ed.) *Expert Systems in the Micro-Electronic Age*, Edinburgh: Edinburgh University Press, reprinted in M. Boden (ed.) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press 1990.

—— (1985a) "Naïve Physics I: Ontology for Liquids",. in J. R. Hobbs and B. Moore (eds) *Formal Theories of the Common Sense World*, Norwood, NJ: Ablex.
—— (1985b) "The Second Naïve Physics Manifesto", in J. R. Hobbs and R. C. Moore (eds) *Formal Theories of the Common Sense World*, Norwood, NJ: Ablex.
Heal, J. (1986) "Replication and Functionalism", in J. Butterfield (ed.) *Language, Mind and Logic*, Cambridge: Cambridge University Press, pp. 135–150, reprinted in M. K. Davies and T. Stone (eds) *Folk Psychology*, Oxford: Blackwell.
—— (1996) "Simulation, Theory, and Content", in P. Carruthers and P. Smith (eds) *Theories of Theories of Mind*, Cambridge: Cambridge University Press.
Heil, J. (1992) *The Nature of True Minds*, Cambridge: Cambridge University Press.
—— (1998) *Philosophy of Mind: A Contemporary Introduction*, London: Routledge.
—— (2003) *From an Ontological Point of View*, New York: Oxford University Press.
—— (ed.) (2004) *Philosophy of Mind: A Guide and Anthology*, New York: Oxford University Press.
Heil, J. and Mele, A. (eds) (1993) *Mental Causation*, Oxford: Oxford University Press.
Hellman, G. (1999) "Reduction(?) to What? Comments on L. Sklar's 'The Reduction(?) of Thermodynamics to Statistical Mechanics'", *Philosophical Studies*, 95: 203–214.
Hempel, C. G. and Oppenheim, P. (1948) "Studies in the Logic of Explanation", *Philosophy of Science*, 15: 135–175.
Heyes, C. and Dickinson, A. (1990) "The Intentionality of Animal Action", *Mind and Language*, 5: 87–104.
—— (1995) "Folk Psychology Won't Go Away: Response to Allen and Bekoff", *Mind and Language*, 10: 329–332.
Hillyard, S. A. (1999) "Electrophysiology, Electric and Magnetic Fields", in R. A. Wilson and F. C. Keil (eds) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.
Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. (1986) "Distributed Representations", in D. E. Rumelhart and J. L. McClelland, and the PDP Research Group *Parallel Distributed Processing*, vol. 1, Cambridge, MA: MIT Press.
Hinton, G. E. and Shallice, T. (1991) "Lesioning an Attractor Network: Investigations of Acquired Dyslexia", *Psychological Review*, 99: 74–95.
Hirschfeld, L. A. and Gelman, S. A. (eds) (1994) *Mapping the Mind: Domain Specificity and Cognition in Culture*, Cambridge: Cambridge University Press.
Hobbs, J. R. and Moore, R. C. (1985) (eds) *Formal Theories of the Common Sense World*, vol. 1, Norwood, NJ: Ablex Publishing Company.
Holland, J. H. (1998) *Emergence: From Chaos to Order*, Reading, MA: Addison-Wesley.
Hollis, M. (1987) *The Cunning of Reason*, Cambridge: Cambridge University Press.
—— (1996) *Reason in Action: Essays in the Philosophy of Social Science*, Cambridge, Cambridge University Press.
Holtzman, S. and Leich, C. (eds) (1981) *Wittgenstein: To Follow a Rule*, London: Routledge and Kegan Paul.
Honderich, T. (1982) "Causes and *if p, even if x, still q*", *Philosophy*, 57: 291–317.
Hooker, C. A. (1981) "Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction", *Dialogue*, 20: 38–59, 201–236, 496–529.
Horgan, T. and Tienson, J. (1989) "Representations Without Rules", *Philosophical Topics*, 27: 147–174.

—— (1992) "Cognitive Systems as Dynamic Systems", *Topoi*, 11: 27–43.
—— (1996) *Connectionism and the Philosophy of Psychology*, Cambridge, MA: MIT Press.
Horgan, T. and Woodward, J. (1985) "Folk Psychology is Here to Stay", *Philosophical Review*, 94: 197–226, reprinted in W. Lycan (ed.) *Mind and Cognition: A Reader*, Oxford: Blackwell, 1990.
Hornsby, J. (1980–81) "Which Mental Events are Physical Events?", *Proceedings of the Aristotelian Society*, 81: 73–92.
—— (1986) "Physicalist Thinking and Conceptions of Behaviour", in P. Pettit and J. McDowell (eds) *Subject, Thought and Context*, Oxford: Oxford University Press.
—— (1997) *Simplemindedness: In Defense of Naïve Naturalism in the Philosophy of Mind*, Cambridge, MA: Harvard University Press.
Horst, S. (1996) *Symbols, Computation and Intentionality: A Critique of the Computational Theory of Mind*, Berkeley, CA: University of California Press.
Horwich, P. (1998) *Meaning*, Oxford: Oxford University Press.
Humphreys, G. W. (1992) *Understanding Vision*, Oxford: Blackwell.
Hurley, S. (1998) *Consciousness in Action*, Cambridge, MA: Harvard University Press.
Hurley, S. and Nudds, M. (2005) *Rational Animals*, Oxford: Oxford University Press.
Huttemann, A. (2004) *What's Wrong with Micro-physicalism?*, London: Routledge.
Jackendorff, R. (1987) *Consciousness and the Computational Mind*, Cambridge, MA: MIT Press.
Jackson, F. (2000) *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford: Clarendon Press.
Jacob, P. (1997) *What Minds Can Do: Intentionality in a Non-Intentional World*, Cambridge: Cambridge University Press.
Jeannerod, M. (1997) *The Cognitive Neuroscience of Action*, Oxford: Blackwell.
Johnson-Laird, P. N. (1988) *The Computer and the Mind: An Introduction to Cognitive Science*, Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N. and Wason, P. C. (1977) "A Theoretical Analysis of Insight into a Reasoning Task", in P. N. Johnson-Laird and P. C. Wason (eds) *Thinking: Readings in Cognitive Science*, Cambridge: Cambridge University Press, pp. 143–157.
Jordan, M. I. and Wolpert, D. M. (2000) "Computational Motor Control", in M. Gazzaniga (ed.) *The New Cognitive Neurosciences*, Cambridge, MA: MIT Press.
Kaiser, M. K., Jonides, J. and Alexander, J. (1986) "Intuitive Reasoning about Abstract and Familiar Physics Problems", *Memory and Cognition*, 14: 308–312.
Kanizsa, G. (1979) *Organization in Vision: Essays on Gestalt Perception*, New York: Praeger.
Karmiloff-Smith, A. (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science*, Cambridge, MA: MIT Press.
—— (1994) "Précis of *Beyond Modularity: A Developmental Perspective on Cognitive Science*", *Behavioral and Brain Sciences*, 17: 693–706.
Kim, J. (1993) *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press.
—— (1996) *Philosophy of Mind*, Oxford: Westview Press.
—— (1998) *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*, Cambridge, MA: MIT Press.
Kirk, R. (1993) " 'The Best Set of Tools': Dennett's Metaphors and the Mind–Body Problem", *Philosophical Quarterly*, 43(172): 335–343.

Kitcher, P. (1984) "In Defense of Intentional Psychology", *Journal of Philosophy*, 81: 89–106.
—— (1988) "Marr's Computational Theory of Vision", *Philosophy of Science*, 55: 1–24.
—— (1990) *Kant's Transcendental Psychology*, New York: Oxford University Press.
Kornblith, H. (ed.) (1985) *Naturalizing Epistemology*, Cambridge, MA: MIT Press.
—— (2002) *Knowledge and Its Place in Nature*, Oxford: Oxford University Press.
Kosslyn, S. M. (1994) *Image and Brain: The Resolution of the Imagery Debate*, Cambridge, MA: MIT Press.
Krebs, J. R. and Kacelnik, A. (1991) "Decision-making", in J. R. Krebs and N. B. Davies (eds) *Behavioral Ecology: An Evolutionary Approach*, Oxford: Blackwell.
Kripke, S. (1982) *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.
Kuczaj, S. A. (1977) "The Acquisition of Regular and Irregular Past Tense Forms", *Journal of Verbal Learning and Verbal Behavior*, 16: 589–600.
Laakso, A. and Cottrell, G. (2000) "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems", *Philosophical Psychology*, 13: 47–76.
Lau, J. (2003) "Externalism about Mental Content", *The Stanford Encyclopedia of Philosophy (Winter 2003 Edition)*, Edward N. Zalta (ed.) url <http://plato.stanford.edu/archives/win2003/entries/content-externalism/>.
Lea, S. E. G. (1984) *Instinct, Environment, and Behaviour*, London: Methuen.
Leahey, T. H. (1992) *A History of Psychology*, 3rd edn, Englewood Cliff, NJ: Prentice Hall.
Ledoux, J. (1996) *The Emotional Brain*, New York: Simon & Schuster.
Lepore, E. and Loewer, B. (1987) "Dual Aspect Semantics", in E. Lepore (ed.) *New Directions in Semantics*, London: Academic Press.
—— (1989) "More on Making Mind Matter", *Philosophical Topics*, 18: 175–191.
Lepore, E. and McLaughlin, B. P. (eds) (1985) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell.
Lepore, E. and Pylyshyn, Z. (eds) (1999) *What Is Cognitive Science?*, Oxford: Blackwell.
Leslie, A. M. (1991) "The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development?", in A. Whiten (ed.) *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, Oxford: Blackwell.
Lewis, D. (1972) "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy*, 50: 249–258, reprinted in N. Block (ed.) *Readings in Philosophy of Psychology*, vol. 1, London: Methuen, 1980. References to the reprinted version.
—— (1973a) *Counterfactuals*, Oxford: Blackwell.
—— (1973b) "Causation", *Journal of Philosophy*, 70: 556–567.
—— (1979) "Counterfactual Dependence and Time's Arrow", *Nous*, 13: 455–476.
—— (1986) *The Plurality of Worlds*, Oxford: Blackwell.
—— (1994) "Reduction of Mind", in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.
Loar, B. (1981) *Mind and Meaning*, Cambridge: Cambridge University Press.
Loewer, B. (1997) "A Guide to Naturalizing Semantics", in B. Hale and C. Wright (eds) *A Companion to the Philosophy of Language*, Oxford: Blackwell.
Loux, M. (ed.) (1979) *The Possible and the Actual: Readings in the Metaphysics of Modality*, Ithaca, NY: Cornell University Press.

Lowe, E. J. (2000) *An Introduction to the Philosophy of Mind*, Cambridge: Cambridge University Press.
Lycan, W. (1987) *Consciousness*, Cambridge, MA: MIT Press.
—— (1988) *Judgment and Justification*, Cambridge: Cambridge University Press.
—— (ed.) (1990/1999) *Mind and Cognition: A Reader*, Oxford: Blackwell.
Macdonald, C. (1995) "Classicism vs. Connectionism", in C. Macdonald and G. Macdonald (eds) (1995b), 3–27.
Macdonald, C. and Macdonald, G. (eds) (1995a) *Philosophy of Psychology: Debates on Psychological Explanation*, Oxford: Basil Blackwell.
—— (eds) (1995b) *Connectionism: Debates on Psychological Explanation*, Oxford: Basil Blackwell.
MacWhinney, B. (2003) "Language Acquisition", in M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press.
McCloskey, M. (1983) "Naïve Theories of Motion", in D. Gentner and A. L. Stevens (eds) *Mental Models*, London: Lawrence Erlbaum Associates.
McCloskey, M., Caramazza, A. and Green, B. (1980) "Curvilinear Motion in the Absence of External Forces: Naive Beliefs about the Motion of Objects", *Science*, 210: 1139–1141.
McDowell, J. (1985) "Functionalism and Anomalous Monism", in E. Lepore and B. P. McLaughlin (eds) *Action and Events*, Oxford: Blackwell.
—— (1987) "In Defence of Modesty", in B. Taylor (ed.) *Michael Dummett: Contributions to Philosophy*, Nijhoff International Philosophy Series, 25, Dordrecht: Nijhoff, 59–80. Reprinted in McDowell (1998).
—— (1994) *Mind and World*, Cambridge, MA: Harvard University Press.
—— (1998a) *Meaning, Knowledge, and Reality*, Cambridge, MA: Harvard University Press.
—— (1998b) *Mind, Value, and Reality*, Cambridge, MA: Harvard University Press.
McGinn, C. (1984) *Wittgenstein on Meaning*, Oxford: Basil Blackwell.
McLeod, P., Plunkett, K. and Rolls, E. T. (1998) *Introduction to Connectionist Modeling of Cognitive Processes*, Oxford: Oxford University Press.
Maloney, J. C. (1989) *The Mundane Matter of the Mental Language*, Cambridge: Cambridge University Press.
Mandler, J. M. (1992) "How to Build a Baby: II, Conceptual Primitives", *Psychological Review*, 99–4: 587–604.
Marcus, G. F. (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive Science*, Cambridge, MA: MIT Press.
Marcus, G. F. *et al.* (1992) *Overregularization in Language Acquisition*, Monographs of the Society for Research in Child Development, 57, Chicago: University of Chicago Press.
Margolis, E. and Laurence, S. (eds) (1999) *Concepts*, Cambridge, MA: MIT Press.
Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, New York: W. H. Freeman and Company.
Matthews, R. J. (1997) "Can Connectionists Explain Systematicity?", *Mind and Language*, 12: 154–177.
Maynard Smith, J. (1982) *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.
Mellars, P. (1996) "Symbolism, Language and the Neanderthal Mind", in P. Mellars and K. Gibson (eds) *Modelling the Early Human Mind: Archaeological and Psychological Perspectives on the Evolution of Human Intelligence*, Cambridge: McDonald Institute Monographs.

Menzies, P. (2001) "Counterfactual Theories of Causation", *The Stanford Encyclopedia of Philosophy (Spring 2001 Edition)*, Edward N. Zalta (ed.) url= <http://plato.stanford.edu/archives/spr2001/entries/causation-counterfactual/>.

Miall, R. C. (2003) "Motor Control, Biological and Theoretical", in M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press.

Miller, A. (1997) "Tacit Knowledge", in B. Hale and C. Wright (eds) *Companion to Philosophy of Language*, Oxford: Blackwell.

Miller, A. and Wright, C. (eds) (2002) *Rule-following and Meaning*, Chesham: Acumen Publishing Ltd.

Millikan, R. G. (1993) "Knowing What I'm Thinking of", *Proceedings of the Aristotelian Society*, supp(67): 91–108.

Milner, B. (1966) "Amnesia Following Operation on the Temporal Lobes", in C. W. M. Whitty and O. L. Zangwill (eds) *Amnesia*, London: Butterworths.

Minsky, M. (1974) "A Framework for Representing Knowledge", in J. Haugeland (ed.) *Mind Design: Philosophy, Psychology, Artificial Intelligence*, Cambridge, MA: MIT Press.

Minsky, M. and Papert, S. (1969) *Perceptrons: An Introduction to Computational Geometry*, Cambridge, MA: MIT Press.

Mithen, S. (1996) *The Prehistory of the Mind*, London: Thames and Hudson.

Morton, A. (1980) *Frames of Mind: Constraints on the Common Sense Conception of the Mental*, Oxford: Clarendon Press.

—— (2003) *The Importance of Being Understood: Folk Psychology as Ethics*, London: Routledge.

Murphy, G. L. (2002) *The Big Book of Concepts*, Cambridge, MA: MIT Press.

Nadel, L. (ed.) (2003) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.

Nagel, E. (1961) *The Structure of Science*, London: Routledge.

Nagel, E. and Newman, J. R. (1958) *Gödel's Proof*, New York: New York University Press.

Nakayama, K. (1999) "Mid-level Vision", in R. A. Wilson and F. C. Keil (eds) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Nakayama, K., He, Z. and Shimojo, S. (1995) "Visual Surface Representation: A Critical Link between Lower-Level and Higher-Level Vision", in D. N. Osherson (ed.) *An Invitation to Cognitive Science: Visual Cognition*, Cambridge, MA: MIT Press.

Nebel, B. (1999) "Frame-Based Systems", in R. A. Wilson and F. C. Keil (eds) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Nelkin, N. (1994) "Patterns", *Mind and Language*, 9(1): 56–87.

Newell, A. and Simon, H. A. (1976) "Computer Science as Empirical Inquiry: Symbols and Search", *Commun. Assoc. Comput. Machinery*, 19: 111–126.

Nichols, S. and Stich, S. (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford: Clarendon Press.

Nichols, S., Stich, S. and Leslie, A. (1995) "Choice Effects and the Ineffectiveness of Simulation", *Mind and Language*, 10(4): 437–445.

Nisbett, R. and Ross, L. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.

O'Connor, T. and Robb, D. (eds) (2003) *Philosophy of Mind: Contemporary Readings*, London: Routledge.

Parker, G. and Maynard Smith, J. (1990) "Optimality Theory in Evolutionary Biology", *Nature*, 348: 27–33.

Patterson, S. (1996) "The Anomalism of Psychology", *Proceedings of the Aristotelian Society*, 96: 37–52.
Peacocke, C. (1986) "Explanation in Computational Psychology: Language, Perception, and Level 1.5", *Mind and Language*, 1(2): 101–123.
—— (1989) "When is a Grammar Psychologically Real?", in A. George (ed.) *Reflections on Chomsky*, Oxford: Blackwell.
—— (1992) *A Study of Concepts*, Cambridge, MA: MIT Press.
—— (1993) "Intuitive Mechanics, Psychological Reality and the Idea of a Material Object", in N. Eilan, R. McCarthy and B. Brewer (eds) *Spatial Representation*, Oxford: Blackwell.
Perner, J. (1996) "Simulation as Explicitation of Predication-Implicit Knowledge about the Mind: Arguments for a Simulation-Theory Mix", in P. Carruthers and P. K. Smith (eds) *Theories of Theories of Mind*, Cambridge: Cambridge University Press.
Pessin, A. and Goldberg, S. (eds) (1996) *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'"*, Armonk, NY: M. E. Sharpe, Inc.
Peterson, M. (1999) "High-Level Vision", in R. A. Wilson and F. C. Keil (eds) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.
Pettit, P. (1991) "Decision Theory and Folk Psychology", in M. Bacharch and S. Hurley (eds) *Foundations of Decision Theory*, Oxford: Blackwell, reprinted in P. Pettit, *Rules, Reasons, and Norms*, Oxford: Oxford University Press, 2002.
Pietrowski, P. and Rey, G. (1995) "When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity", *British Journal for the Philosophy of Science*, 46: 81–110.
Pinker, S. (1994) *The Language Instinct*, New York: HarperCollins.
—— (1997) *How the Mind Works*, New York: Norton.
Pinker, S. and Prince, A. (1988) "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition", *Cognition*, 28: 73–193.
Pitcher, G. (1977) *Berkeley*, London: Routledge and Kegan Paul.
Plantinga, A. (1974) *The Nature of Necessity*, Oxford: Oxford University Press.
Plunkett, K. (1995) "Connectionist Approaches to Language Acquisition", in P. Fletcher and B. MacWhinney (eds) *The Handbook of Child Language*, Oxford: Blackwell.
Plunkett, K. and Marchman, V. (1993) "From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets", *Cognition*, 48: 21–69.
Pollard, P. and Evans, J. (1987) "Content and Context Effects in Reasoning", *American Journal of Psychology*, 100–1: 41–60.
Pomerantz, J. (2003) "Perception, Overview", in L. Nadel (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.
Posner, M. (ed.) (1989) *Foundations of Cognitive Science*, Cambridge, MA: MIT Press.
Posner, M. I. and Raichle, M. E. (1994) *Images of Mind*, New York: Scientific American Library.
Poundstone, W. (1985) *The Recursive Universe*, New York: William Morrow and Company.
Price, H. (1996) *Time's Arrow and Archimedes' Point*, New York: Oxford University Press.
Prince, A. and Pinker, S. (1988) "Rules and Connections in Human Language", *Trends*

in Neurosciences, 11: 195–202, reprinted in R. Cummins and D. D. Cummins (eds) *Minds, Brains, and Computers*, New York: Oxford University Press, 1999.

Prinz, J. (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis*, Cambridge, MA: MIT Press.

Proffitt, D. R. (1999) "Naïve Physics", in R. A. Wilson and F. C. Keil (eds) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Putnam, H. (1960) "Minds and Machines", in S. Hook (ed.) *Dimensions of Mind*, New York: New York University Press.

—— (1962) "The Analytic and the Synthetic", in H. Feigl and G. Maxwell (eds) *Scientific Explanation, Space, and Time, Minnesota Studies in the Philosophy of Science*, vol. 3, Minneapolis: University of Minnesota Press, reprinted in H. Putnam (1975) *Mind, Language and Reality*, Cambridge: Cambridge University Press.

—— (1965) "How Not to Talk about Meaning: Comments on J. J. C. Smart", in R. S. Cohen and M. R. Wartofsky (eds) *Boston Studies in the Philosophy of Science*, vol. 2, New York: Humanities Press.

—— (1967) "The Nature of Mental States", reprinted in *Mind, Language and Reality* (1975).

—— (1975) "The Meaning of 'Meaning'", in K. Gunderson (ed.) *Language, Mind, and Knowledge*, Minneapolis: University of Minnesota Press, pp. 131–193. Also in H. Putnam (1975) *Mind, Language and Reality*, Cambridge: Cambridge University Press. References are to Putnam, *Mind, Language and Reality*.

—— (1983) "Philosophers and Human Understanding", in A. F. Heath (ed.) *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, Oxford: Clarendon Press.

Pylyshyn, Z. W. (1980) "Cognition and Computation: Issues in the Foundations of Cognitive Science", *Behavioral and Brain Sciences*, 3–1: 154–169.

—— (1981) "The Imagery Debate: Analogue Media Versus Tacit Knowledge", *Psychological Review*, 88: 16–45.

—— (1984) *Computation and Cognition: Towards a Foundation for Cognitive Science,* Cambridge, MA: MIT Press.

Quine, W. V. O. (1951) "Two Dogmas of Empiricism", *The Philosophical Review*, 60: 20–43.

—— (1969) "Epistemology Naturalized", in *Ontological Relativity*, New York: Columbia University Press, reprinted in H. Kornblith (ed.) *Naturalizing Epistemology*, Cambridge, MA: MIT Press.

—— (1974) *The Roots of Reference*, Cambridge, MA: Harvard University Press.

Ramachandran, V. S. (1988) "Perceiving Shape from Shading", *Scientific American*, 256(6): 76–83.

Ramsey, W., Stich, S. and Garon, J. (1991) "Connectionism, Eliminativism and the Future of Folk Psychology", in W. Ramsey, D. Rumelhart and S. Stich (eds) *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum.

Ramsey, W., Rumelhart, D. and Stich, S. (eds) (1991) *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum.

Rey, G. (1995) "A Not 'Merely Empirical' Argument for a Language of Thought", *Philosophical Perspectives*, 9: 201–222.

—— (1997) *Contemporary Philosophy of Mind*, Oxford: Basil Blackwell.

Rips, L. J. (1983) "Cognitive Processes in Propositional Reasoning", *Psychological Review*, 90: 38–71.

Rock, I. (1983) *The Logic of Perception*, Cambridge, MA: MIT Press.

Rogers, R. (1971) *Mathematical Logic and Formalized Theories*, Amsterdam: North-Holland Publishing Co.
Rorty, R. (1970) "In Defence of Eliminative Materialism", *Review of Metaphysics*, 24: 112–121.
Rosenthal, D. (ed.) (1991) *The Nature of Mind*, New York: Oxford University Press.
Ruben, D.-H. (1990) "Singular Explanation and the Social Sciences", in P. French, T. Uehling, Jr. and H. Wettstein (eds) *Midwest Studies in Philosophy: The Philosophy of the Human Sciences*, Notre Dame, IN: University of Notre Dame Press, vol XV.
—— (ed.) (1993) *Explanation*, Oxford: Oxford University Press.
—— (1994) "A Counterfactual Theory of Causal Explanation", *Nous*, 28: 465–481.
Rumelhart, D. E. (1989) "The Architecture of Mind: A Connectionist Approach", in M. Posner (ed.) *Foundations of Cognitive Science*, Cambridge, MA: MIT Press.
Rumelhart, D. E., Hinton, G. and Williams, R. (1986) "Learning Representations by Back-propagating Errors", *Nature*, 323–9: 533–536.
Rumelhart, D. E., McClelland, J. L. and the PDP Research Group (1986) *Parallel Distributed Processing*, vol. 1, Cambridge, MA: MIT Press.
Ryle, G. (1949) *The Concept of Mind*, London: Penguin.
Saffran, E. M. (1982) "Neuropsychological Approaches to the Study of Language", *British Journal of Psychology*, 73: 317–337.
Samuels, R., Stich, S. and Tremoulet, P. D. (1999) "Rethinking Rationality: From Bleak Implications to Darwinian Modules", in E. Lepore and Z. Pylyshyn (eds) *What Is Cognitive Science?*, Oxford: Blackwell.
Sanford, D. H. (1989) *If P, then Q. Conditionals and the Foundations of Reasoning*, London: Routledge.
Schaffner, K. F. (1967) "Approaches to Reduction", *Philosophy of Science*, 34: 137–147.
Schick, F. (1991) *Understanding Action: An Essay on Reasons*, New York: Cambridge University Press.
—— (1997) *Making Choices: A Recasting of Decision Theory*, New York: Cambridge University Press.
Schiffer, S. (1991) "Ceteris Paribus Laws", *Mind*, 100: 1–17.
—— (1994) "Thought and Language", in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind*, Oxford: Blackwell.
Schiffman, H. R. (2001) *Sensation and Perception: An Integrated Approach*, 5th edn, New York: John Wiley and Sons Inc.
—— (2003) "Psychophysics", in S. F. Davis (ed.) *Handbook of Research Methods in Experimental Psychology*, Malden, MA: Blackwell.
Scholl, B. J. and Leslie, A. M. (1999) "Modularity: Development and Theory of Mind", *Mind and Language*, 14(1): 131–153.
Searle, J. (1990a) "Is the Brain's Mind a Computer Program?", *Scientific American*, 262: 26–31.
—— (1990b) "Consciousness, Inversion, and Cognitive Science", *Behavioral and Brain Sciences*, 13(4): 585–596.
Segal, G. (1991) "Defence of Reasonable Individualism", *Mind*, 100: 485–494.
—— (1996) "The Modularity of Theory of Mind", in P. Carruthers and P. K. Smith (eds) *Theories of Theories of Mind*, Cambridge: Cambridge University Press.
—— (2000) *A Slim Book about Narrow Content*, Cambridge, MA: MIT Press.
Sellars, W. (1969) "Language as Thought and as Communication", *Philosophy and Phenomenological Research*, 29: 506–527.

Shallice, T. (1988) *From Neuropsychology to Mental Structure*, Cambridge: Cambridge University Press.
Shallice, T. and Warrington, E. K. (1980) "Single and Multiple Component Central Dyslexic Syndromes", in M. Coltheart, K. E. Patterson and J. C. Marshall (eds) *Deep Dyslexia*, London: Routledge and Kegan Paul.
Shastri, L. (2003) "Structured Connectionist Model", in M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press.
Shepard, R. N. (1994) "Perceptual-Cognitive Universals as Reflections of the World", *Psychonomic Bulletin and Review*, 1: 2–28, reprinted 2001 in *Behavioral and Brain Sciences*, 24(4): 581–601.
Shepard, R. N. and Cooper, L. A. (1982) *Mental Images and Their Transformations*, Cambridge, MA: MIT Press.
Shoemaker, S. (1984) *Identity, Cause and Mind*, Cambridge: Cambridge University Press.
Siegelmann, H. and Sontag, E. (1991) "Turing Computability with Neural Nets", *Applied Mathematics Letters*, 4(6): 77–80.
Sklar, L. (1993) *Physics and Chance. Philosophical Issues in the Foundations of Statistical Mechanics*, Cambridge: Cambridge University Press.
—— (1999) "The Reduction(?) of Thermodynamics to Statistical Mechanics", *Philosophical Studies*, 95: 187–202.
Skyrms, B. (1996) *The Evolution of the Social Contract*, Cambridge: Cambridge University Press.
Smith, A. D. (1992) "Modest Reductions and the Unity of Science", in D. Charles and K. Lennon (eds) *Reduction, Explanation, and Realism*, Oxford: Clarendon Press.
—— (1993) "Non-Reductive Physicalism?", in H. Robinson (ed.) *Objections to Physicalism*, Oxford: Clarendon Press.
Smith, N. H. (2002) *Reading McDowell*, London: Routledge.
Smolensky, P. (1988) "On the Proper Treatment of Connectionism", *Behavioral and Brain Sciences*, 11–1: 1–23.
—— (1991) "Connectionism, Constituency, and the Language of Thought", in B. Loewer and G. Rey (eds) *Meaning in Mind: Fodor and his Critics*, Oxford: Basil Blackwell.
—— (1995) "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture", in C. Macdonald and G. Macdonald (eds) *Connectionism: Debates on Psychological Explanation*, vol. 2, Oxford: Basil Blackwell.
Sosa, E. and Tooley, M. (eds) (1993) *Causation*, Oxford: Oxford University Press.
Stalnaker, R. (2003) *Ways A World Might Be: Metaphysical and Anti-Metaphysical Essays*, Oxford: Oxford University Press.
Stein, B. E., Wallace, M. T. and Stanford, T. R. (1998) "The Use of Single Neuron Electrophysiology in Cognitive Science", in W. Bechtel and A. Graham (eds) *A Companion to Cognitive Science*, Malden, MA: Blackwell.
Stein, E. (1996) *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*, New York: Oxford University Press.
Sterelny, K. (1990) *The Representational Theory of Mind*, Oxford: Basil Blackwell.
Stern, D. N. (1985) *The Interpersonal World of the Infant*, New York: Basic Books.
Stich, S. P. (1978) "Beliefs and Subdoxastic States", *Philosophy of Science*, 45: 499–518.
—— (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, MA: MIT Press.

—— (1990) *The Fragmentation of Reason*, Cambridge, MA: MIT Press.
—— (1996) *Deconstructing the Mind*, New York: Oxford University Press.
Stich, S. P. and Nichols, S. (1992) "Folk Psychology: Simulation or Tacit Theory", *Mind and Language*, 7(1): 35–71.
—— (1995) "Second Thoughts on Simulation", in M. K. Davies and T. Stone (eds) *Mental Simulation*, Oxford: Blackwell.
Stich, S. and Warfield, T. (1995) "Do Connectionist Minds have Beliefs? – A Reply to Clark and Smolensky", in C. Macdonald and G. Macdonald (eds) *Connectionism: Debates on Psychological Explanation*, vol. 2, Oxford: Blackwell.
—— (eds) (2003) *Blackwell Guide to the Philosophy of Mind*, New York: Oxford University Press.
Taylor, C. (1964) *The Explanation of Behavior*, London: Routledge and Kegan Paul.
Thagard, P. (1996) *Mind: Introduction to Cognitive Science*, Cambridge, MA: MIT Press.
Tiffany, E. C. (2001) "The Rational Character of Belief and the Argument for Mental Anomalism", *Philosophical Studies*, 103–3: 285–314.
Tinbergen, N. (1951) *The Study of Instinct*, Oxford: Clarendon Press.
—— (1973) *The Animal in its World*, Cambridge, MA: Harvard University Press.
Todd, P. M. and Gigerenzer, G. (2000) "Simple Heuristics that Make Us Smart", *Behavioral and Brain Sciences*, 23(5): 727–742.
Travis, C. (1994) "On Constraints of Generality", *Proceedings of the Aristotelian Society*, 94: 165–188.
—— (2001) *Unshadowed Thought*, Cambridge, MA: Harvard University Press.
Tulving, E. (1972) "Episodic and Semantic Memory", in E. Tulving and W. Donaldson (eds) *Organisation of Memory*, New York: Academic Press.
Tversky, A. and Kahneman, D. (1971) "Belief in the Law of Small Numbers", *Psychological Bulletin*, 76: 105–110.
—— (1974) "Judgment Under Uncertainty: Heuristics and Biases", *Science*, 185: 1124–1131, reprinted in P. K. Moser (ed.) *Rationality in Action: Contemporary Approaches*, Cambridge: Cambridge University Press, 1990.
—— (1983) "Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment", *Psychological Review*, 91: 293–315.
Tye, M. (1990) "A Representational Theory of Pains and their Phenomenal Character", in J. Tomberlin (ed.) *Philosophical Perspectives*, vol. 9, Atascadero, CA: Ridgeview Publishing Co.
—— (1991) *The Imagery Debate*, Cambridge, MA: MIT Press.
Ullman, S. (1996) *High-level Vision*, Cambridge, MA: MIT Press.
van Gelder, T. (1990) "Compositionality: A Connectionist Variation on a Classical Theme", *Cognitive Science*, 14: 355–384.
van Gulick, R. "Functionalism, Information and Content", *Nature and System* 2: 139–162. References are to the reprinted version in W. G. Lycan (ed.) *Mind and Cognition: A Reader*, Oxford: Blackwell, 1990.
Walker, S. (1983) *Animal Thought*, London: Routledge and Kegan Paul.
Wason, P. and Johnson-Laird, P. (1972) *Psychology of Reasoning: Structure and Content*, Cambridge, MA: Harvard University Press.
Weiskrantz, L. (ed.) (1988) *Thought Without Words*, Oxford: Clarendon Press.
Weiss, B. (2002) *Michael Dummett*, Princeton, NJ: Princeton University Press.
Wilson, R. A. (1994) "Wide Computationalism", *Mind*, 103: 351–372.
Wilson, R. A. and Keil, F. C. (eds) (1999) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Wimmer, H. and Perner, J. (1983) "Beliefs about Beliefs: Representing and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception", *Cognition*, 13: 103–128.

Wittgenstein, L. (1953) *Philosophical Investigations*, trans. G. E. M Anscombe, Oxford: Basil Blackwell.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Wraga, M. and Kosslyn, S. M. (2003) "Imagery", in L. Nadel (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group.

Wright, C. (1981) "Rule-following, Objectivity and the Theory of Meaning", in S. Holtzman and C. Leich (eds) *Wittgenstein*, London: Routledge and Kegan Paul.

—— (1986) "Theories of Meaning and Speakers' Knowledge", in S. Shanker (ed.) *Philosophy in Britain Today*, London: Croom Helm.

Yalowitz, S. (1997) "Rationality and the Argument for Anomalous Monism?", *Philosophical Studies*, 87–3: 235–258.

Index