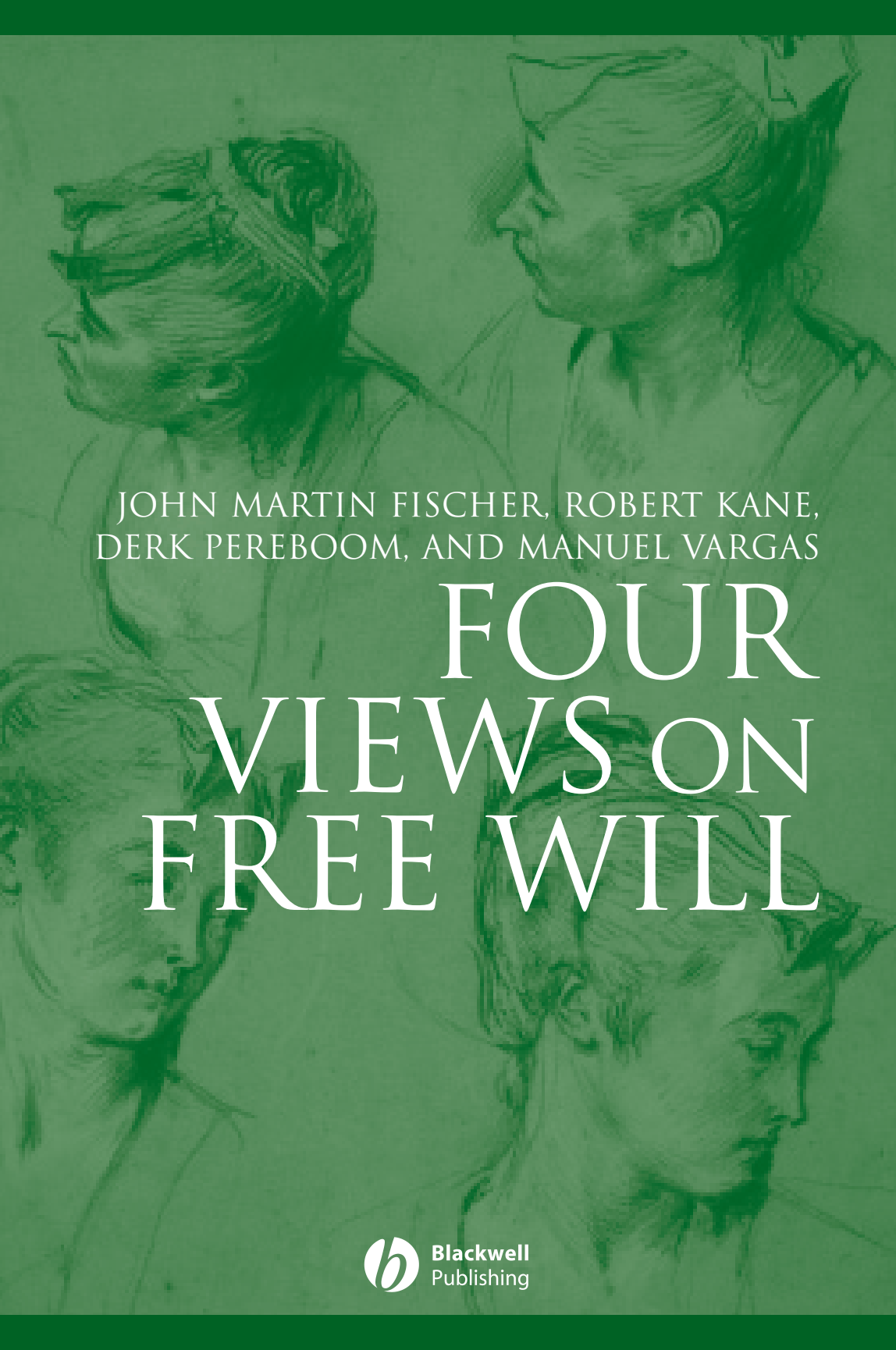

JOHN MARTIN FISCHER, ROBERT KANE,
DERK PEREBOOM, AND MANUEL VARGAS

# FOUR VIEWS ON FREE WILL

Blackwell
Publishing

# Four Views on Free Will

**Great Debates in Philosophy**
*Series Editor: Ernest Sosa*

Dialogue has always been a powerful means of philosophical exploration and exposition. By presenting important current issues in philosophy in the form of a debate, this series attempts to capture the flavor of philosophical argument and to convey the excitement generated by the exchange of ideas. Each author contributes a major, original essay. When these essays have been completed, the authors are each given the opportunity to respond to the opposing view.

Personal Identity
*Sydney Shoemaker and Richard Swinburne*

Consciousness and Causality
*D. M. Armstrong and Norman Malcolm*

Critical Theory
*David Couzens Hoy and Thomas McCarthy*

Moral Relativism and Moral Objectivity
*Gilbert Harman and Judith Jarvis Thomson*

Atheism and Theism, Second Edition
*J. J. C. Smart and J. J. Haldane*

Three Methods of Ethics
*Marcia W. Baron, Philip Pettit, and Michael Slote*

Epistemic Justification
*Laurence BonJour and Ernest Sosa*

Four Views on Free Will
*John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas*

# Four Views on Free Will

John Martin Fischer,
Robert Kane,
Derk Pereboom,
and
Manuel Vargas

BLACKWELL PUBLISHING
350 Main Street, Malden, MA 02148–5020, USA
9600 Garsington Road, Oxford OX4 2DQ, UK
550 Swanston Street, Carlton, Victoria 3053, Australia



First published 2007 by Blackwell Publishing Ltd

1 2007

*Library of Congress Cataloging-in-Publication Data*

Four views on free will / John Martin Fischer . . . [et al.].
p. cm. — (Great debates in philosophy)
Includes bibliographical references (p. ) and index.
ISBN-13: 978-1-4051-3485-9 (hardback)
ISBN-13: 978-1-4051-3486-6 (pbk.)
1. Free will and determinism. I. Fischer, John Martin, 1952–

BJ1461.F68 2007
123′.5—dc22

2006026270

A catalogue record for this title is available from the British Library.

Set in 10 on 12.5 pt Caslon
by SNP Best-set Typesetter Ltd, Hong Kong
Printed and bound in Singapore
by Markono Print Media Pte Ltd

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on
Blackwell Publishing, visit our website:
www.blackwellpublishing.com

# *Contents*

# *Notes on Contributors*

**John Martin Fischer** is Distinguished Professor in the Department of Philosophy at the University of California, Riverside, where he holds a UC President's Chair. He is the author of *The Metaphysics of Free Will: An Essay on Control* (Blackwell, 1994); and, with Mark Ravizza, S.J., *Responsibility and Control: A Theory of Moral Responsibility* (1998). His collection of essays, *My Way: Essays on Moral Responsibility*, was published in 2006.

**Robert Kane** is University Distinguished Teaching Professor of Philosophy at the University of Texas at Austin. He is the author of seven books and over sixty articles on the philosophy of mind and action, ethics, the theory of value and philosophy of religion, including *Free Will and Values* (1985), *Through the Moral Maze: Searching for Absolute Values in a Pluralistic World* (1994), *The Significance of Free Will* (1996), *A Contemporary Introduction to Free Will* (2005), and a lecture series on audio and video tape entitled *The Quest for Meaning: Values, Ethics, and the Modern Experience*. He is editor *of The Oxford Handbook of Free Will* (2002) and of *Free Will* (Blackwell, 2002). The recipient of fifteen major teaching awards at the University of Texas, he was named an inaugural member of the University's Academy of Distinguished Teachers in 1995.

**Derk Pereboom** is Professor of Philosophy at the University of Vermont. He will join the Sage School of Philosophy at Cornell University in 2007. His book *Living Without Free Will* appeared in 2001, and he has published articles on free will, philosophy of mind, history of modern philosophy, and philosophy of religion.

**Manuel Vargas** is Associate Professor of Philosophy at the University of San Francisco. The author of various articles in ethics, philosophy of action, and Latin American philosophy, he has been awarded the American Philosophical Association's Prize in Latin American Thought (2004), and the N.E.H. Chair in the Humanities at the University of San Francisco (2005–2006).

# *Acknowledgments*

JMF: I'd like to thank Neal A. Tognazzini and John T. Maier for their helpful comments on my essay.

DP: Thanks from me to Manuel Vargas, Robert Kane, John Fischer, Seth Shabo, David Christensen, and Sarah Adler.

MV: Thanks to the other authors, and also to Eddy Nahmias, Shaun Nichols, and Dan Speak for very helpful feedback on my contributions to this book. Thanks to Neal Tognazzini for the index. Thanks also to Stephanie for lots of things.

RK: Thanks to the other three authors for their cogent comments, and to Manuel for his diligent organizational efforts.

# *A Brief Introduction to Some Terms and Concepts*

## Basic Terms: Free Will, Moral Responsibility, and Determinism

Perhaps the three most important concepts in philosophical work on free will are *free will*, *moral responsibility*, and *determinism*.

The notion of freedom at stake in philosophical discussions is usually distinguished from a variety of other freedom concepts, including things like religious and political freedom. Usually, **free will** is also treated as distinct from several other concepts associated with human agency, such as autonomy and authenticity. As we will see in the chapters that follow, there are many different ways of thinking about the nature of free will, and there are serious disagreements about what would constitute an adequate theory of free will. Much of the tradition has taken "free will" to be a kind of power or ability to make decisions of the sort for which one can be morally responsible, but philosophers have also sometimes thought that free will might be required for a range of other things, including moral value, originality, and self-governance. Two other claims often made about free will are hotly disputed among philosophers; and authors of this volume will take different sides on these claims. One is the claim that free will requires "alternative possibilities" or the power to do otherwise, and the other is the claim that free will requires that we are the "ultimate sources" of our free actions or the ultimate sources of our wills to perform free actions.

Important to many discussions of free will is the idea of **moral responsibility**. In the context of discussions of free will, moral responsibility is

often understood as a kind of status connected to judgments and/or practices of moral praise and blame. This meaning is distinct from another, perhaps more commonly used sense of responsibility: responsibilities as obligations (for example, when we talk about what responsibilities a parent has to a child). There are important connections between responsibility of the sort concerned with praise and blame and responsibility of the sort connected with obligations. However, philosophers writing on free will and moral responsibility are typically concerned with the former and not the latter.

**Determinism** is a third concept that is often important for philosophical discussions of free will. For present purposes, we can treat determinism as the thesis that at any time (at least right up to the very end) the universe has exactly one physically possible future. Something is deterministic if it has only one physically possible outcome.

It is important to bear in mind that a definition of determinism is just that – a characterization of what things would have to be like *if* things were deterministic. It does not follow that the universe is actually deterministic. Compare: "A creature is a gryphon if it has the hindquarters of a lion and the head and claws of an eagle." Nothing about the definition of gryphon shows that there are such creatures in our universe. It simply tells us something about what sorts of things would count as gryphons. Similarly, to offer a definition of determinism does not show that the universe is deterministic. It only defines a term, and we may find that the term never properly applies to the world we live in.

When discussing these issues it is natural to wonder whether the world is deterministic. Most physicists and philosophers think that the answer is no, but the technical issues are extremely complex. Nevertheless, if we accept that the universe isn't deterministic there are still good reasons to think about the compatibility of free will and determinism. First, it could turn out that future physicists conclude that the universe is deterministic, contrary to the contemporary consensus about at least quantum mechanics. It is notoriously difficult to predict how future science will turn out, and it might be useful to have an answer to the question in advance of the scientific issues getting sorted out. Second, even if the universe were not fully deterministic, determinism might hold locally (either as a matter of how local spacetime is constructed, or as a matter of how the physics for non-quantum physical objects operates). Third, we could be interested in whether free will is compatible with a broadly scientific picture of the universe. Since some aspects of the universe seem deterministic and others do not, we might ask if free will is compatible with determinism as a first step to answering the more general question of whether free will is compatible with a broadly scientific picture of the universe.

## Philosophical Options on the Free Will Problem

One particularly important issue for contemporary philosophers thinking about free will is whether we could have free will in a deterministic universe. Call this issue – whether free will could exist if the universe were deterministic – **the compatibility issue.** There is a long-standing tradition of dividing up the conceptual terrain in light of the main answers to the compatibility issue. Traditionally, **incompatibilists** are those who think that free will is incompatible with the world being deterministic. **Compatibilists**, conveniently enough, are those hold that free will is compatible with the universe being deterministic.

It is important to recognize that the compatibility issue is distinct from the issue of whether we have free will. You could be an incompatibilist, and maintain that we have do have free will. Or you might be an incompatibilist and think that we lack free will. (You could even think that irrespective of how the compatibility issue is settled, there are threats to free will apart from determinism.)

In the philosophical literature, **libertarianism** is the view that we have free will and that free will is incompatible with determinism. "Libertarianism" as it is used in the context of free will is distinct from libertarianism in political philosophy. (Indeed, "libertarianism" in the free will sense is the original meaning – it was only later appropriated as the label for a view in political philosophy.) One might be a libertarian in both political and free will senses, but you can be a libertarian about free will without being a libertarian in political philosophy. And, perhaps, you could also be a political libertarian without being a free will libertarian (although many political libertarians seem to also be free will libertarians).

Following Derk Pereboom, we will label as "**hard incompatibilism**" any view that holds that (1) incompatibilism is true and (2) we lack free will. Historically, most hard incompatibilists were what William James called **hard determinists**. (Indeed, Pereboom's coining of the term "hard incompatibilism" reflects James' older and narrower terminology.) Hard determinists think we lack free will *because the world is deterministic.* Contemporary hard determinists are few and far between. What is more common are views that hold that we have no free will irrespective of whether or not the world is deterministic, and views that hold that although freedom might be not be conceptually incompatible with determinism (or indeterminism, for that matter), we simply do not have it.

To summarize, then: A traditional way of dividing up the terrain concerns answers to the compatibility issue. The two main approaches are incompatibilism and compatibilism. We have been considering the incompatibilist fork, where the two main species of incompatibilism are libertarianism and hard

incompatibilism. Both forms of incompatibilism have further species we have not discussed in this brief introduction.

The remaining fork of the compatibility debate is **compatibilism**. There are many varieties of compatibilism. Some compatibilists have emphasized a particular understanding of "can," others have emphasized a kind of identification with one's motives or values, and others emphasizing the role of responsiveness to reasons. One influential variation, however, is the view that holds that responsibility is compatible with determinism, combined with agnosticism about whether free will understood in some particular way might not be compatible with determinism. This view is ***semicompatibilism***, and its most prominent defender is John Martin Fischer.

Lastly, there are views that do not neatly fit the traditional taxonomy of incompatibilism and compatibilism. One such class of views is **revisionism**. The core idea of revisionism is that the picture of free will and moral responsibility embedded in commonsense is in need of revision, but not abandonment. That is, the revisionist holds that the correct account of free will and moral responsibility will depart from commonsense. As is the case with libertarianism, hard incompatibilism, and compatibilism, this view can take a variety of more specific forms.

For a different way to think about the relationship between the various views, see the grid below.

| | Is commonsense thinking about free will and moral responsibility basically correct? | Is free will compatible with determinism? | Is moral responsibility compatible with determinism? | Do we have free will? |
|---|---|---|---|---|
| Libertarianism | Yes | No | No | Yes |
| Compatibilism | Yes | Yes (although semicompatibilists may say "no") | Yes | Yes |
| Hard Incompatibilism | No | No | No | No |
| Revisionism | No | Yes, but only with revision to our self-image | Yes | Yes (or close enough) |

# 1
# *Libertarianism*

*Robert Kane*

## 1 Determinism and the Garden of Forking Paths

The problem of free will has arisen in history whenever people have been led to suspect that their actions might be determined or necessitated by factors unknown to them and beyond their control. That is why doctrines of *determinism* or *necessity* have been so important in the history of debates about free will.

Doctrines of determinism have taken many historical forms. People have wondered at various times whether their actions might be determined by Fate or by God, by the laws of physics or the laws of logic, by heredity or environment, by unconscious motives or hidden controllers, psychological or social conditioning, and so on. But there is a core idea running through all historical doctrines of determinism that shows why they are all a threat to free will. All doctrines of determinism – whether they are fatalistic, theological, physical, biological, psychological or social – imply that, given the past and the laws of nature at any given time, there is only one possible future. Whatever happens is therefore inevitable or necessary (it cannot but occur), given the past and the laws.

To see why many persons have believed there is a conflict between free will and determinism, so conceived, consider what free will requires. We believe we have free will when we view ourselves as agents capable of influencing the world in various ways. Open alternatives seem to lie before is. We reason and deliberate among them and choose. We feel (1) it is "up to us" what we choose and how we act; and this means we could have chosen or acted otherwise. As Aristotle said, "when acting is 'up to us,' so is not acting." This "up-to-us-ness" also suggests that (2) the ultimate sources of our actions lie in us and not outside us in factors beyond our control.

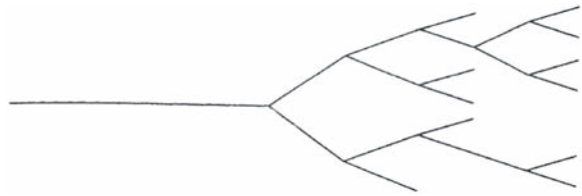

**Figure 1** Garden of Forking Paths

To illustrate, suppose Jane has just graduated from law school and she has a choice between joining a law firm in Chicago or a different firm in New York. If Jane believes her choice is a *free* choice (made "of her own free will"), she must believe both options are "open" to her while she is deliberating. She could choose either one. (If she did not believe this, what would be the point of deliberating?) But that means she believes there is more than one possible path into the future available to her and it is "up to her" which of these paths will be taken. Such a picture of an open future with forking paths – a garden of forking paths, it has been called – is essential to our understanding of free will.

This picture of different possible paths into the future is also essential, I believe, to what it means to be a person and to live a human life.

One can see why determinism would threaten this picture. If determinism is true, it seems there would not be more than one possible path into the future available to Jane, but only one. It would not be (1) "up to" her what she chose from an array of alternative possibilities, since only one alternative would be possible. It also seems that, if determinism were true, the (2) sources or origins of her actions would not be in Jane herself but in something else outside her control that determined her choice (such as the decrees of fate, the foreordaining acts of God, her heredity and upbringing or social conditioning).

A second way to illustrate why many people believe there is a conflict between free will and determinism is to reflect on the idea of *responsibility*. Free will is also intimately related to notions of accountability, blameworthiness and praiseworthiness for actions.

Suppose a young man is on trial for an assault and robbery in which his victim was beaten to death. Let us say we attend his trial and listen to the evidence in the courtroom. At first, our thoughts of the young man are filled with anger and resentment. His crime was heinous. But as we listen daily to how he came to have the mean character and perverse motives he did have – a sad story of parental neglect, child abuse, sexual abuse, bad role models –

some of our resentment against the young man is shifted over to the parents and others who abused and mistreated him. We begin to feel angry with them as well as with him. (Note how natural this reaction is.) Yet we aren't quite ready to shift all of the blame away from the young man himself. We wonder whether some residual responsibility may not belong to him. Our questions become: To what extent is *he* responsible for becoming the sort of person he now is? Was it *all* a question of bad parenting, societal neglect, social conditioning, and the like, or did he have any role to play in it?

These are crucial questions about free will and they are questions about what may be called the young man's ultimate responsibility**.** We know that parenting and society, genetic make-up and upbringing, have an influence on what we become and what we are. But were these influences entirely *determining* or did they "leave anything over" for us to be responsible for? That is what we want to know about the young man. The question of whether he is merely a victim of bad circumstances or has some responsibility for being what he is – the question, that is, of whether he became the person he is *of his own free will* – seems to depend on whether these other factors were or were not *entirely* determining.

Those who are convinced that there is a conflict between free will and determinism, for these and other reasons, are called *incompatibilists* about free will. They believe free will and determinism are incompatible. If incompatibilists also believe that an incompatibilist free will exists, so that determinism is false, they are called *libertarians* about free will.

## 2 Modern Challenges to Libertarian Free Will

I will be defending the libertarian view of free will in this volume. We libertarians typically believe that a free will that is incompatible with determinism is required for us to be truly morally responsible for our actions, so that genuine moral responsibility, as well as free will, is incompatible with determinism. Genuine free will, we believe, could not exist in a world that was *completely* determined by Fate or God, or the laws of physics or logic, or heredity and environment, psychological or social conditioning, and so on. In writings over the past twenty-five years, I have argued that this libertarian view represents the traditional idea of free will that has been in dispute for centuries when philosophers have discussed "the problem of free will and determinism." Moreover, I think this libertarian view is the one many ordinary persons have in mind when they intuitively believe there is some kind of conflict between free will and determinism.

Yet this traditional libertarian conception of free will has been under attack by many modern thinkers, philosophers and scientists alike, who have come

to believe that such an idea of free will, though it may still be held by many ordinary people, is outmoded and incoherent and that it has no place in the modern scientific picture of the world. A goal of this essay is therefore to consider this modern attack on the traditional libertarian view of free will and to ask how, and whether, it can be answered. Much is at stake, it seems to me, in knowing whether we do or do not have a freedom of the will of the ultimate kind that libertarians defend. The modern attack on it has two parts.

*Part 1* The first prong of the modern attack on libertarian free will comes from *compatibilists*, who argue that, despite appearances to the contrary, determinism does not really conflict with free will at all. Compatibilists argue that all the freedoms we recognize and desire in ordinary life – e.g., freedoms from coercion or compulsion, from physical restraint, from addictions and political oppression, for example – are really compatible with determinism. Even if the world should turn out to be entirely deterministic, compatibilists argue, there would still be a big difference between persons who are free from constraints on their freedom of action and will (constraints such as coercion, compulsion, addiction and oppression) and persons who are not free from these constraints; and people would prefer to be free from such constraints on their freedom rather than not, *even in a determined world*. Thus, according to compatibilists, esoteric questions about whether determinism is true or not – in the physical or psychological sciences – are irrelevant to *the freedoms we really care about* in everyday life. All the varieties of free will "worth wanting" (as a modern compatibilist, Daniel Dennett, has put it) do not require the falsity of determinism for us to possess them, as the traditional libertarian view of free will suggests.

This doctrine of *compatibilism* has an ancient lineage. It was held by the Stoics and perhaps also by Aristotle in ancient times, according to many scholars. But compatibilism about free will and determinism has become especially popular in modern times. Influential philosophers of the modern era, such as Thomas Hobbes, John Locke, David Hume and John Stuart Mill, were all compatibilists. They saw compatibilism as a way of reconciling ordinary experience of being free with modern scientific views about the universe and human beings; and compatibilism continues to be popular among philosophers and scientists today for similar reasons, as you will see from later essays of this volume. (John Martin Fischer defends a version of compatibilism, known as semicompatibilism, in the second essay of this volume.) If compatibilists are right, we can have *both* free will and determinism; and we need not worry that increasing scientific knowledge about nature and human beings will somehow undermine our ordinary convictions that we are free and responsible agents.

*Part 2* The second prong of the modern attack on libertarian free will goes a step further. Recall that the first prong says that libertarian free will is *unnecessary* because we can have all the freedoms worth wanting, even if determinism should be true. The second prong goes further, arguing that libertarian free will itself is *impossible* or *unintelligible* and has no place in the modern scientific picture of the world. Such an ultimate freedom is not something we could have anyway, say its critics. Those who take this line note that defenders of libertarian free will have often invoked obscure and mysterious forms of agency or causation to defend the libertarian view. In order to explain how free actions can escape the clutches of physical causes and laws of nature (so that free actions will not be determined by physical laws), libertarians have posited transempirical power centers, immaterial egos, noumenal selves outside of space and time, unmoved movers, uncaused causes and other unusual forms of agency or causation – thereby inviting charges of obscurity or mystery against their view. Even some of the greatest modern defenders of libertarianism, such as Immanuel Kant, have argued that we need to believe in libertarian free will to make sense of morality and genuine responsibility, but we can never completely understand such a free will in theoretical and scientific terms.

The problem that provokes this widespread skepticism about the existence of libertarian free will has to do with an ancient dilemma: If free will is not compatible with *determinism*, as libertarians contend, free will does not seem to be compatible with *indeterminism* either (the opposite of determinism). Events that are undetermined, such as quantum jumps in atoms, happen merely by chance. So if free actions were undetermined, as libertarians claim, it seems that they too would happen by chance. But how can chance events be free and responsible actions? Suppose a choice was the result of a quantum jump or other undetermined event in a person's brain. Would this amount to a free and responsible choice? Undetermined effects in the brain or body would be unpredictable and impulsive – like the sudden emergence of a thought or the uncontrolled jerking of an arm – quite the opposite of what we take free and responsible actions to be. It seems that undetermined events in the brain or body would occur *spontaneously* and would be more likely to *undermine* our freedom rather than *enhance* it.

This two-pronged modern attack on the traditional libertarian view of free will has had a powerful impact on modern thought. To answer it, libertarians must show (i) that free will really *is* incompatible with *determinism* (call this "The Compatibility Problem"). But they must also show (ii) that a libertarian free will requiring *indeterminism* can be made intelligible and how, if at all, such a free will can be reconciled with modern scientific views of the cosmos and of human beings (call this "The Intelligibility Problem"). I will be

addressing both these problems in this chapter, beginning with the first, or "Compatibility Problem."

## 3 Is Free Will Incompatible with Determinism?: The Consequence Argument

The popularity of compatibilism among modern philosophers and scientists means that libertarians who believe free will is incompatible with determinism can no longer merely rely on intuitions about "forking paths" into the future to support their view that determinism conflicts with free will (as in section 1). These intuitions must be backed up with arguments that show *why* free will must be incompatible with determinism. To meet this challenge, libertarians have proposed new arguments for incompatibilism in modern philosophy; and we will begin by considering the most widely discussed of these new arguments for the incompatibility of free will and determinism.

This important argument is called the "Consequence Argument" and it is stated informally as follows by one of its proponents, Peter van Inwagen:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore the consequences of these things (including our own acts) are not up to us. (From *An Essay on Free Will*, Oxford: Clarendon Press, 1983, p. 16)

To say it is not "up to us" what "went on before we were born," or "what the laws of nature are," is to say that there is nothing we can now do to change the past or alter the laws of nature (it is beyond our control). We can thus spell out this Consequence Argument in the following steps:

(1) There is nothing we can now do to change the past.
(2) There is nothing we can now do to change the laws of nature.
(3) There is nothing we can now do to change the past and the laws of nature.
(4) If determinism is true, our present actions are necessary consequences of the past and the laws of nature. (That is, it *must* be the case that, given the past and the laws of nature, our present actions occur.)
(5) Therefore, there is nothing we can now do to change the fact that our present actions occur.

In other words, we *cannot now do otherwise* than we actually do. Since this argument can be applied to any agents and actions at any times, we can infer

from it that *if determinism is true, no one can ever do otherwise*; and if free will requires the power to do otherwise than we actually do (as in the image of forking paths), then no one would have free will.

Defenders of the Consequence Argument, such as van Inwagen, think the first two premises are undeniable. We cannot now change the past (1) or the laws of nature (2). Step 3 states what appears to be a simple consequence of premises 1 and 2: If you can't change the past or the laws, then you can't change the conjunction of both of them. Premise 4 simply spells out what is implied by determinism. Some philosophers have questioned one or another of the first three steps of this argument. But most criticisms have focused on step 5. Step 5 follows from 3 and 4 by virtue of the following inference: If (3) there is nothing we can now do to change the past and laws of nature and (4) our present actions are necessary consequences of the past and laws, then (5) there is nothing we can now do to change the fact that our present actions occur. This inference is an instance of the following principle:

> (TP) If there is nothing anyone can do to change X, and if Y is a necessary consequence of X (if it must be that, if X occurs, Y occurs), then there is nothing anyone can do to change Y.

TP has been called a "Transfer of Powerlessness Principle" for it says in effect that if you are powerless to change something X, and something else Y is necessarily going to occur if X does, then you are also powerless to change Y. This makes sense. If we can't do anything to prevent X from occurring and Y cannot but occur if X does, then how could we do anything to prevent Y from occurring? Consider an example. Suppose the sun is going to explode in AD 2050 and there is nothing anyone can now do to change the fact that the sun will explode in AD 2050. Assume also that necessarily (given the laws of nature), if the sun explodes in AD 2050, all life on earth will end in AD 2050. If both these claims are true, it seems obvious that there is nothing anyone can now do to change the fact that all life on earth will end in 2050. Here is another example. If there is nothing anyone can now do to change the laws of nature, and the laws of nature entail that nothing goes faster than the speed of light, then there is nothing anyone can now do to change the fact that nothing goes faster than the speed of light.

But, despite the initial plausibility of this Transfer of Powerlessness Principle, critics of the Consequence Argument have challenged it. Everything depends, they say, on how you interpret the expression "There is nothing anyone can do to change . . ." Talking about what persons "can" (and "cannot") do is talking about their *powers*; and the notion of power is one of the most difficult in metaphysics, as John Locke pointed out three centuries ago. For example, many *compatibilists* interpret what it means to

say that persons "can" or "have the power" to do things in the following way. They say

> "You can (or you have the power to do) something."

simply means

> "If you wanted (or tried) to do it, you would do it."

I can jump over this fence means I would jump over it, if I wanted to or tried to. If someone challenged my power to do it, the challenger would say "I don't think you would manage to jump it, *even if* you wanted or tried."

Now the interesting thing about this compatibilist interpretation of "can" and "power," is that, if it is correct, the Consequence Argument would fail. For on this interpretation, to say we can now change the past or the laws would mean that

> "*If* we now wanted or tried to change the past or the laws, we would change them."

And this is false. No persons would change the past or the laws of nature, *even if* they wanted or tried to, because no one has the power to do it. But when we turn to ordinary actions like jumping over a fence, things are different. If you can jump over a fence that is in your path, it may well be true that you *would* jump over it, *if* you wanted to or tried, because jumping over fences is something you *are* capable of doing.

In other words, on the analysis of "can" or "power" that many compatibilists favor, the *premises* of the Consequence Argument come out *true* (you would *not* have changed the past or the laws, even if we wanted or tried to, because you are not capable of it). But the *conclusion* of the Consequence Argument comes out *false* (you would have jumped the fence, *if* you wanted or tried to, because jumping fences of this height is something you *are* capable of doing). Since the Consequence Argument would have true premises and a false conclusion on this analysis of "can," it would be an invalid argument. What has happened to make it fail? The answer is that the transfer principle TP has failed. Your powerlessness to change the past and laws of nature does not *transfer* to your powerlessness to jump the fence. For you are *not* able to change the past and laws, but you *are* able to jump the fence – at least in this compatibilist sense that ("you would do it, *if* you wanted or tried to."

But why should we accept this "hypothetical" compatibilist account of "can" or "power" ("you would do it, *if* you wanted or tried to")? Defenders of the Consequence Argument, such as van Inwagen, do not accept this

hypothetical account of "can" or "power"; nor do most libertarians. They would respond to the preceding compatibilist argument as follows:

> "So the Consequence Argument fails on your compatibilist analysis of 'can' or 'power.' But that should not surprise us. For your compatibilist analysis was rigged in the first place to make freedom compatible with determinism. On your analysis, persons can jump the fence even though their doing so here and now is impossible, given the past and the laws of nature. That is not what we libertarians mean by 'can' in the Consequence Argument. We mean it is possible that you do it *here and now*, *given all the facts that presently obtain*. If your analysis allows you to say that persons can do otherwise, even though they can't change the past and the laws of nature and even though their actions are a necessary consequence of the past and the laws of nature, *then something must be wrong with your compatibilist analysis*. What use is a power or ability to do something, if it cannot be *exercised* in the existing circumstances here and now? To us libertarians, the premises and rules of the Consequence Argument are far more plausible than any compatibilist analysis of 'can.'"

At this point, arguments over the Consequence Argument tend to reach an impasse. Incompatibilist defenders of the argument claim that compatibilist critics are begging the question by interpreting "can" in the Consequence Argument in a way that is compatible with determinism. But compatibilists respond by saying that defenders of the Consequence Argument are begging the question themselves by assuming that "can" in the argument has an *incompatibilist* meaning rather than a compatibilist one.

## 4 Ultimate Responsibility

As a result of this impasse, philosophical debates have multiplied about just what "can" and "power" (and related expressions, such as "could have done otherwise") really mean. We cannot follow all these complex debates here. But I do not think it matters. For I believe disagreements over the meaning of "can" and "power" are symptoms of a deeper problem in discussions about free will and determinism. The problem is that focusing on "alternative possibilities" (or "forking paths" into the future) or the "power to do otherwise" *alone*, as the Consequence Argument does, is *too thin a basis* on which to rest the case for the incompatibility of free will and determinism. One must look beyond debates about "can," "power," "ability," and "could have done otherwise" to make the case for the incompatibility of free will and determinism.

Fortunately, there is another place to look for reasons why free will might conflict with determinism. Recall that in section 1, I suggested that there were *two* reasons why people thought determinism must rule out free will.

One was the requirement of (1) alternative possibilities we have been considering: Free will seems to require that *open alternatives* or *alternative possibilities* lie before us – a garden of forking paths – and it is "up to us" which of these alternatives we choose. (Call this condition "AP" for "alternative possibilities"). But there was a second condition mentioned that has also historically fueled incompatibilist intuitions: (2) Free will also seems to require that the *sources* or *origins* of our actions lie "in us" rather than in something else (such as the decrees of fate, the foreordaining acts of God, or antecedent causes and laws of nature) outside us and beyond our control.

I call this second requirement for free will the condition of Ultimate Responsibility (or UR, for short); and I think it is even more important to free will debates than AP, or alternative possibilities. The basic idea of UR is this: *To be ultimately responsible for an action, an agent must be responsible for anything that is a sufficient cause or motive for the action's occurring.* If, for example, a choice issues from, and can be sufficiently explained by, an agent's character and motives (together with background conditions), then to be *ultimately* responsible for the choice, the agent must be at least in part responsible by virtue of choices or actions voluntarily performed in the past for having the character and motives he or she now has. Compare Aristotle's claim that if a man is responsible for the wicked acts that flow from his character, he must at some time in the past have been responsible for forming the wicked character from which these acts flow.

This condition of Ultimate Responsibility, or UR, does not require that we could have done otherwise (AP) for *every* act done of our own free wills. But it does require that we could have done otherwise with respect to *some* acts in our past life histories by which we formed our present characters. I call these earlier acts by which we formed our present characters "self-forming actions," or SFAs.

To see why such self-forming acts are important for free will, consider a well-known example about Martin Luther offered by Daniel Dennett. When Martin Luther finally broke with the Church in Rome, initiating the Protestant Reformation, he said "Here I stand, I can do no other." Now Dennett asks us to suppose that at the moment Luther made this stand, he was literally right. Given his character and motives, Luther *could* not then and there *have done otherwise*. Does this mean Luther was not morally responsible, not subject to praise or blame, for his act, or that he was not acting of his own free will? Dennett says "not at all." In saying "I can do not other," Luther was not disowning responsibility for his act, according to Dennett, but taking full responsibility for acting of his own free will. So the ability to do otherwise ("could have done otherwise") or AP, says Dennett, is not required for moral responsibility or free will.

Now Dennett is a compatibilist, as noted earlier, and he is using this Luther example to defend compatibilism of free will and determinism by

suggesting that free will and moral responsibility do not even require the power to do otherwise or alternative possibilities (AP). Note that, if this were true, the Consequence Argument would be undermined. We would not have to get into complex debates about what "could have done otherwise" means, since free will and moral responsibility would not require alternative possibilities (AP) or "could have done otherwise" in the first place.

But, now, if we look at Dennett's Luther example from the point of view of the condition of Ultimate Responsibility or UR, rather than simply in terms of AP, there is an answer that can be given to Dennett. We can grant that Luther could have been responsible for this act, even though he could *not* have done otherwise *then and there* and even if his act was determined. But this would be so, if UR is required, only to the extent that Luther was responsible for his present motives and character by virtue of some *earlier* struggles and self-forming actions (SFAs) that brought him to this point in his life where he could do no other. Those who know Luther's biography know the inner struggles and turmoil he endured getting to that point in his life. Often we act from a will already formed, but it is "our own free will" by virtue of the fact that *we* formed it by other choices or actions in the past (SFAs) for which we *could* have done otherwise. If this were not so, *there is nothing we could have ever done to make ourselves different than we are* – a consequence, I believe, that is incompatible with our being (at least to some degree) ultimately responsible (UR) for what we are. So SFAs are only a subset of those acts in life for which we are ultimately responsible and which are done "of our own free will." But if *none* of the acts in our lifetimes were self-forming in this way, we would not be *ultimately* responsible for anything we did.

If the case for incompatibility of free will and determinism cannot be made by reference to AP alone, it can be made if UR is added. So, I suggest, the often-neglected condition of ultimate responsibility or UR should be moved to center stage in free will debates. If agents must be responsible to some degree for anything that is a *sufficient cause* or *motive* for their actions (as UR requires), then an impossible infinite regress of past actions would be required unless some actions in the agent's life history (SFAs) did not have either sufficient causes or motives (and hence were undetermined). Therein lies the connection between UR and determinism. If we must have formed our present wills (our characters and motives) by earlier voluntary choices or actions, then UR would require that if any of these earlier choices or actions *also* had sufficient causes or motives when we performed *them*, then we must have also been responsible for those earlier sufficient causes or motives by virtue of forming them by *still earlier* voluntary choices or actions, and so on backwards indefinitely into our past. Eventually we would come to infancy or to a time before our birth when we could not have formed our own wills.

The only way to stop this regress is to suppose that *some* acts in our life histories must lack *sufficient* causes altogether, and hence must be undetermined, if we are to be the ultimate sources or grounds of, and hence ultimately responsible for, our own wills. These regress-stopping acts would be the "self-forming acts" or SFAs that are required by UR sometime in our lives, if we are to have free will. Note, as a result, that UR makes explicit something that is often hidden in free will debates, namely that *free will* – as opposed to mere *freedom of action* – is about the forming and shaping of character and motives which are the *sources* or *origins* of praiseworthy or blameworthy, virtuous or vicious, actions. Free *will* (in contrast to mere free *action*) is about *self-formation*. If persons are responsible for the wicked (or noble, shameful, heroic, generous, treacherous, kind or cruel) acts that flow from their wills (characters and motives), they must at some point be responsible for forming the *wills* from which these acts flow.

## 5 Ultimate Responsibility and Alternative Possibilities

Another thing to note about this argument for the incompatibility of free will and determinism from UR is that – unlike the Consequence Argument – the argument from UR does not mention the condition of *alternative possibilities* or AP at all. The argument from UR says that, if agents must be responsible to some degree for anything that is a *sufficient cause or motive* for their actions (as UR requires), then an impossible infinite regress of past actions would be required, *unless* some actions in the agent's life history (SFAs) did not have either sufficient causes or motives and hence were undetermined. The argument from UR thus focuses on the sources or origins of what we actually do rather than on the power to do otherwise.

When one argues about the incompatibility of free will and determinism from alternative possibilities or AP (as in the Consequence Argument), the focus is on notions of "necessity," "possibility," "power," "ability," "can," and "could have done otherwise." By contrast, the argument from UR focuses on a different set of concerns about the "sources," "grounds," "reasons," and "explanations" of our wills, characters, and purposes. Where did our motives and purposes come from, who produced them, who is responsible for them? Was it *we* ourselves who are responsible for forming our characters and purposes, or someone or something else – God, fate, heredity and environment, nature or upbringing, society or culture, behavioral engineers or hidden controllers? Therein lies the core of the traditional problem of free will.

But does this mean that alternative possibilities or AP have nothing to do with free will? It might seem so, if one can argue directly for the incompatibility of free will and determinism from UR without mentioning alternative

possibilities. But then what would become of the garden of "forking paths" if alternative possibilities or AP are not required? Well, fortunately it turns out that AP and the garden of forking paths *is* relevant for free will after all. For, it can be shown that *UR entails AP* for at least *some* free actions. Why this is so is not obvious, but understanding it is also crucial, I believe, to fully understand the nature of free will.

To understand the connection between AP and UR, alternative possibilities and ultimate responsibility, we must first note that having alternative possibilities for one's action – though it may be necessary for free will – is not *sufficient* for free will, *even if* the alternative possibilities should also be *undetermined*. This can be shown by noting that there are examples in which agents may have alternative possibilities *and* their actions are undetermined, *and yet the agents lack free will*. This sounds strange. But it is important for understanding free will to understand how it could be. I call examples that show this "Austin-style examples" after the British philosopher J. L. Austin, who suggested the first example of this kind in free will debates.

Here are three easily understood "Austin-style examples" that I will refer to in later arguments. The first example is Austin's own. (i) He imagined that he had to hole a three-foot putt to win a golf match but, owing to a nervous twitch in his arm, he misses the putt. The other two examples are mine. (ii) An assassin is trying to kill the prime minister with a high-powered rifle when, owing to a nervous twitch in his arm, he misses and kills the minister's aide instead. (iii) I am standing in front of a coffee machine intending to press the button for coffee without cream when, owing to a brain cross, I accidentally press the button for coffee with cream. Now notice that in each of these examples, we can suppose, as Austin suggests, that an element of genuine chance or indeterminism is involved. Perhaps the nervous twitches or brain crosses are brought about by actual undetermined quantum jumps in our nervous systems. We can thus imagine that Austin's holing the putt is a genuinely undetermined event. He might miss the putt by chance and, in the example, does miss it by chance. (Likewise, the assassin might hit the wrong target by chance and I might press the wrong button by chance.)

Now Austin asked the following question about his example: Can we say in these circumstances that "he (Austin) could have done otherwise" than miss the putt? Did he have alternative possibilities? Austin's answer is that we can indeed say he could have done otherwise than miss it. For he was a good putter. He had made many similar putts of this short length in the past (he had the capacity and opportunity to make it). But even more important, since the outcome of this putt was genuinely *undetermined*, he might well have succeeded in holing the putt, as he was trying to do.

But this means we have an action (missing the putt) that is (i) *undetermined* and (ii) such that the agent could have done otherwise. (In other words, we

have indeterminism *plus* alternative possibilities or AP.) Yet missing the putt is not something that we regard as *freely* done in any normal sense of the term because it is not under the agent's voluntary control. Austin missed the putt all right; and he *could* have holed it – he could have done otherwise. But he did not miss it *voluntarily* and *freely*. He did not choose to miss it. The same is true of the assassin's failing to hit the prime minister and killing the aide and my accidentally pressing the wrong button on the coffee machine. Both of us could have done otherwise (the assassin could have hit his target and I could have pressed the right button) because our actions were undetermined and they might have gone the other way. Yet the assassin did not miss his target voluntarily and as a result of his own free choice; and I did not press the wrong button voluntarily and as a result of my own free choice.

One might be tempted to think that these three occurrences (missing the putt, killing the aide, pressing the wrong button) are not *actions* at all in such circumstances because they are undetermined and happen by accident. But Austin correctly warns against drawing such a conclusion. Missing the putt, he says, was clearly something he *did,* even though it was not what he *wanted* or *chose* to do. Similarly, killing the aide was something the assassin did, though unintentionally; and pressing the wrong button was something I did, even if only by accident or inadvertently. Austin's point is that many of the things we do *by accident* or *mistake*, *unintentionally* or *inadvertently*, are nonetheless things we *do*. We may sometimes be absolved of responsibility for doing them (though not always, as in the case of the assassin). But it is for *doing* them that we are absolved of responsibility; and this can be true even if the accidents or mistakes are genuinely undetermined.

But now we can draw a further conclusion from these Austin-style examples (the conclusion we were looking for) that Austin himself did not consider. These examples also show that alternative possibilities *plus* indeterminism are not sufficient for *free will* (even if they should be necessary). To see why, suppose that God created a world in which there is a lot of indeterminism of the kind that occurs in Austin-style examples. Chance plays a significant role in this world, in human affairs as well as in nature. People set out to do things and often succeed, but sometimes they fail in the manner of Austin-style examples. They set out to kill prime ministers, hole putts, press buttons on coffee machines, thread needles, punch computer keys, scale walls, and so on – usually succeeding, but sometimes failing by mistake or accident in ways that are undetermined.

Now imagine further that in this world all actions of all agents, whether they succeed in their purposes or not, are such that their reasons, motives and purposes for trying to act as they do are always predetermined or pre-set by God. Whether the assassin misses the prime minister or not, his intent to kill the prime minister in the first place is predetermined by God. Whether

or not Austin misses his putt, his wanting and trying to make it rather than miss are preordained by God. Whether I press the button for coffee without cream, my wanting to do so because of my dislike of cream is predetermined by God; and so it is for all persons and all of their actions in this imagined world. Their reasons, motives and purposes for acting as they do are always predetermined by God.

I would argue that persons in such a world lack *free will*, even though it is often the case that they can do otherwise (and thus have alternative possibilities) in a way that is undetermined. The reason is that they can do otherwise, but only in the limited Austin-style way – by mistake or accident, unwillingly or unintentionally. What they cannot do in any sense is *will* otherwise than they do; for all of their reasons, motives and purposes have been pre-set by God. We may say that the wills of persons in this world are always already "set one way" before and when they act, so that if they do otherwise, it will not be "in accordance with their wills."

The possibility of such worlds shows in a striking way why, to have free will, it is necessary not only to be the ultimate source of one's *actions*, but also to be the ultimate source of one's *will* to perform the actions. It would not be enough for free will to be unhindered in the pursuit of one's motives and purposes, if all of one's motives and purposes were created by someone or something else (God or fate or whatever) as in the above-imagined world. Even one's motives or purposes for wanting to change one's motives or purposes would be created by someone or something else in such a world.

Now UR captures this additional requirement of being the ultimate source of one's *will* that is lacking in this imagined world. For UR says that we must be responsible by virtue of our voluntary actions for anything that is a sufficient cause or a sufficient *motive* (or *reason)* for our acting as we do. We have a sufficient motive or reason for doing something, when our will is "set one way" on doing it before and when we act – as the assassin's will is set on killing the prime minister. Among the available things he might do, only one of them (killing the prime minister) would be voluntary and intentional. Anything else he might do (miss the prime minister, kill the aide) would be done only by accident or mistake, unintentionally or unwillingly.

UR says that if you have a sufficient motive for doing something in this sense – if your will is "set one way" on doing it rather than anything else available to you – then to be ultimately responsible for your *will*, you must be to some degree responsible by virtue of past voluntary acts for your will's being set the way it is. This is significant because, when we look to the responsibility of the assassin for what he did, we look to his evil motives and intentions. They are the source of his guilt, whether he succeeds in killing the prime minister or fails and kills the aide instead. Luther too, we assumed,

had a sufficient motive for his final affirmation, "Here I stand." Yet, we said that if Luther's will was firmly set one way by the time he made his affirmation, this would not count against his being ultimately responsible, *so long as he was responsible for his will's being set that way*. That is what UR requires.

But now it looks like we have another regress on our hands. If it should turn out that our wills were already set one way when we performed the earlier voluntary actions *by* which we set our present wills, then UR would require that we must have been responsible by virtue of still earlier voluntary actions for our wills' being set the way they were at that earlier time, and so on backwards indefinitely. But, once again, this is only a *potential* regress. Just as the regress discussed earlier could be stopped by assuming that some actions in an agent's history lacked *sufficient causes*, so this regress can be stopped by supposing that some actions in an agent's past also lacked *sufficient motives*. Actions lacking sufficient motives would be actions in which the agents' wills were not already set one way *before* they performed them. Rather, the agents would set their wills one way or another in the performance of the actions themselves.

We may call such actions in which agents "set their wills" in one way or another in the performance of the actions themselves "will-setting" actions. Will-setting actions occur, for example, when agents make choices or decisions between two or more competing options and do not settle on which of the options they want more, all things considered, until the moment of choice or decision itself. They thus "set" their wills in one way or the other in the act of choosing itself.

The need for such will-setting actions tells us something further about free will. When we wonder about whether agents have freedom of will (rather than wondering only about whether they have freedom of action), what interests us is not merely whether they could have done otherwise, even if the doing otherwise is undetermined, but whether they could have done otherwise *voluntarily* (or *willingly*), *intentionally,* and *rationally*. Or, more generally, we are interested in whether they could have acted in *more than one way* voluntarily, intentionally, and rationally, rather than (as in the Austin-style examples) only in one way voluntarily, intentionally, and rationally and in other ways merely by accident or mistake, unintentionally or irrationally. ("Voluntarily" means here "in accordance with one's will"; "intentionally" means "knowingly" and "on purpose" and "rationally," means "having good reasons for acting and acting for those reasons.")

We might call these requirements of *more-than-one-way* (or plural) voluntariness, rationality, and intentionality, "*plurality conditions*" for free will. Such conditions seem to be deeply embedded in our intuitions about free choice and action. Most of us naturally assume that freedom and responsibility would be deficient if it were always the case that we could only do otherwise

by accident or mistake, unintentionally, or involuntarily. Free will seems to require that if we acted voluntarily, intentionally, and rationally, we could also have done otherwise voluntarily, intentionally, and rationally. But *why* do we assume this so readily; and why are these plurality conditions so deeply embedded in our intuitions about free will?

The argument of the previous section from UR provides the clue. If (i) *free will* requires (ii) *ultimate responsibility* for our wills as well as for our actions, then it requires (iii) *will-setting* actions at some points in our lives; and will-setting actions require (iv) the *plurality conditions*, the ability to act in more than one way voluntarily, intentionally and rationally. To see why will-setting actions require the plurality conditions, consider a variation on the assassin example that would make his choice to kill the prime minister a will-setting one. Suppose that just before pulling the trigger, the assassin has doubts about his mission. Pangs of conscience arise in him and a genuine inner struggle ensues about whether or not to go through with the killing. The assassin now has more than one motivationally significant option before his mind. So his will is no longer clearly set one way; and he will only resolve the issue one way or the other by consciously deciding and thereby setting his will in one direction or the other. Unlike the original assassin example, neither outcome in this case (where he is conscience-stricken and has to decide one way or the other) would be a mere accident or mistake; either resolution would be a voluntary and intentional decision to go through with the killing or to stop. Such a will-setting action would therefore be voluntary, intentional, and rational whichever way it goes and so it would satisfy the plurality conditions.

So we have the following chain of inferences: (1) *free will* entails (2) *ultimate responsibility* [UR] for our wills as well as for our actions, which entails (3) *will-setting* actions at some points in our lives, which in turn entails that some of our actions must satisfy (4) the *plurality conditions*. But if actions satisfy the plurality conditions and the agents could have done otherwise voluntarily, intentionally, and rationally, then the agents could have done otherwise; and so they had (5) *alternative possibilities*. Therein lies the connection between UR and AP. If free will requires ultimate responsibility in the sense of UR, then at least *some* actions in our life histories ("will-setting actions") must be such that we could have done otherwise with respect to them. Note, however, that this argument from (1) free will to (5) alternative possibilities (AP) is not direct. It goes *through* (2) ultimate responsibility (UR), (3) will-setting and (4) plurality; and UR is the key to it, since it is UR that implies will-setting and plurality. If we are to be ultimately responsible for our own *wills*, some of our actions must be such that we could have done otherwise, *because* some of them must have been such that we could have done otherwise voluntarily, intentionally, and rationally.

UR thus entails both indeterminism *and* alternative possibilities or AP. But it entails them by different argumentative routes. Two separate regresses are involved. (I call this the "dual regress of free will.") The first regress begins with the requirement (of UR) that agents must be responsible by virtue of past voluntary actions for anything that is a *sufficient cause* of their actions. Stopping this regress requires that if agents are to have free will, some actions in their life histories must be *undetermined* (must lack sufficient causes). The second regress begins with the requirement that agents be responsible by virtue of past voluntary actions for anything that is a sufficient *motive* for their actions. Stopping this regress requires that some actions in an agent's life history must be will-setting (so they do not have sufficient motives already set) and hence must satisfy the plurality conditions and hence AP.

The first of these two regresses results from the requirement that we be ultimate sources of our *actions*, the second from the requirement that we be ultimate sources of our *wills* (to perform those actions). If the second requirement were not added, we might have a world in which all the will-setting was done by someone or something other than the agents themselves, as in the imagined world in which all the will-setting was done by God. Agents in such a world might be unhindered in the pursuit of their purposes or ends, but it would never be "up to them" what *purposes* or ends they pursued. To have free will therefore is to be the ultimate designer of one's own purposes or ends or goals. And if we are to be the ultimate designers of our own purposes or ends, there must be *some* actions in our life histories that are will-setting, plural voluntary *and* undetermined by someone or something else.

These undetermined, will-setting actions are the "*self-forming actions*," or SFAs required by UR mentioned earlier. They would be the actions in our lives by which we ultimately *form* our character and motives and make ourselves into the kinds of persons we are.

## 6 The Intelligibility Problem: Is Libertarian Free Will Possible?

Can we make sense of a free will that requires Ultimate Responsibility of the kind described in the previous section? Can we really be the ultimate designers of our own ends and purposes? There are many skeptics about free will who think not. They argue that being the *ultimate* source of one's will and actions is an incoherent and impossible ideal, since it would require us to be "prime movers unmoved" or "uncaused causes of ourselves" – "the best self-contradiction that has been conceived so far," as Friedrich Nietzsche put it. Ultimate Responsibility or UR requires that there be some acts in our life

histories that do not have sufficient causes or motives. But how could acts having neither sufficient causes nor motives be free and responsible actions?

This question brings us to the second part of the modern attack on libertarian free will. It is one thing to offer arguments showing that free will is not compatible with determinism (and hence to address the "Compatibility Problem"). It is quite another thing to answer charges that an incompatibilist free will requiring ultimate responsibility is *intelligible* or *possible* and can be reconciled with modern scientific views of human beings. This is the "Intelligibility Problem" about libertarian free will; and it is in many ways even more difficult than the Compatibility Problem.

The culprit in the case of the Intelligibility Problem is not determinism, but *indeterminism*. For the Intelligibility Problem is related to an ancient dilemma noted earlier: if free will is not compatible with determinism, it does not seem to be compatible with indeterminism either. The arguments to show this have been made since ancient times. An undetermined or chance event, it is said, occurs spontaneously and is not controlled by anything, hence not controlled by the agent. To cite an example mentioned earlier, if a choice occurred by virtue of a quantum jump or other undetermined event in one's brain it would seem a fluke or accident rather than a responsible choice. Such undetermined events occurring in our brains or bodies would not seem to enhance our freedom and control over our actions, but rather diminish our freedom and control.

Or we could put the Intelligibility Problem in another way that goes a little deeper. If my free choice is really undetermined, that means I could have made a different choice *given exactly the same past* right up to the moment when I did choose. That is what indeterminism and probability mean: given exactly the same past, different outcomes ("forking paths") are possible. Imagine, for example, that John had been deliberating about where to spend his vacation, in Hawaii or Colorado, and after much thought and deliberation, had decided he preferred Hawaii and chose it. If the choice was undetermined, then exactly the same deliberation, the same thought processes, the same beliefs, desires, and other motives – not a sliver of difference – that led up to John's favoring and choosing Hawaii over Colorado, might by chance have issued in his choosing Colorado instead. That is very strange. If such a thing happened it would seem a fluke or accident, like that quantum jump in the brain just mentioned, not a rational choice. Since John had come to favor Hawaii and was about to choose it, when by chance he chose Colorado, he might well wonder what went wrong and perhaps consult a neurologist.

One may at first think that there must be some way around the conclusion that if a choice is undetermined (like John's choice just described), then the agent must have been able to choose otherwise "*given exactly the same past*." But in fact there is no easy way around this conclusion. For indeterminism,

which is the denial of determinism, *does* mean "different possible futures, given the same past." In the diagram of forking paths in section 1, the single line going back into the past is just that: a single line indicating "same past"; while the multiple lines going into the future represent "different possible futures." By contrast, determinism means only one line into the future. If John is really free to choose different options at any time during his deliberation, and his choice is not determined, then he must be able to choose *either* path (Hawaii or Colorado), given the *same* past up to the moment when he chooses.

You can't cheat here and say "If the past had been just *a tiny bit* different, then John might have sensibly and rationally chosen differently (chosen Colorado instead)." *Determinists* and *compatibilists* can say this. For they insist that John might have *sensibly* and *rationally* chosen otherwise only if the past had been different in some way (however small the difference). For example, if John had had a few different desires and beliefs or had reasoned a little differently, he might have come to favor Colorado and chosen it instead of Hawaii. But persons who believe free choices cannot be determined (as libertarians do) must say John may have chosen different possible futures, given the same entire past, including his psychological and physical history up to the moment he did choose. And this does seem to make his choosing otherwise (choosing Colorado) arbitrary and irrational in the same circumstances in which he actually came to favor Hawaii and chose it. You can see why many people have argued that undetermined free choices, of the kind libertarians demand, would be "arbitrary," "capricious," "random," "irrational," "uncontrolled," and "inexplicable," and not really free and responsible choices at all.

Defenders of libertarian free will, according to their critics, have a dismal record of answering such charges. Realizing that free will cannot merely be indeterminism or chance, libertarians have appealed to various unusual forms of agency or causation to make up the difference. For example, Immanuel Kant said we cannot explain free will in scientific and psychological terms, even though we require it for belief in morality. To account for free will, we have to appeal to the agency of what Kant called a "noumenal self" outside space and time that could not be studied in scientific terms. Many other respectable philosophers continue to believe that only some sort of appeal to mind/body dualism, of the kind associated with Descartes, can make sense of free will. Science might tell us there was indeterminacy or a place for causal gaps in the brain, but a non-material self or soul, or what Nobel physiologist John Eccles calls a "transempirical power center," would have to fill the causal gaps left by physical causes by intervening in the natural order. The most popular appeal among libertarians today is to a special kind of *agent- or immanent causation* that cannot be explained in terms of the ordinary modes

of causation in terms of events familiar to the sciences. Free and responsible actions are not determined by prior events, according to this "agent-causation" view, but neither do free actions occur merely by chance. They are caused by the *agent (*a substance) in a way that transcends and cannot be explained in terms of ordinary modes of causation by events or states of affairs involving the agent.

I call these familiar libertarian strategies for making sense of free will "extra factor" strategies. The general idea behind all such strategies is not hard to understand: Since indeterminism means that an agent might act one way or in a different way, given exactly the same past, which would seem to include all the *same prior mental and physical events*, some "extra" kind of causation or agency must be postulated over and above the natural flow of events to account for the agent's going one way rather than another. In short, some additional factor must be involved to "tip the balance." It is this line of thought that has led libertarians through the centuries to postulate extra factors, such as immaterial causes, noumenal selves, transempirical power centers, non-event agent causes, prime movers unmoved, and so on, to explain free choices. And these postulates have in turn brought down on libertarians charges of obscurantism or mystery or "panicky metaphysics" from their critics.

Now it may be that some extra factors of the kinds just mentioned (or some others) *are* necessary to make sense of libertarian free will. Most libertarians today believe, for example, that some notion of "agent-causation" or causation by a substance that does not consist in causation in terms of events or states of affairs involving the agent, is required to make sense of free will. And this agent-causation view is ably defended by a number of recent philosophers, including Roderick Chisholm, Timothy O'Connor, Randolph Clarke, William Rowe, and others. But I happen to agree with other libertarians about free will, such as Peter van Inwagen and Carl Ginet, that "extra factor" strategies – including agent-causation theories – do not solve the problems about indeterminism they are suppose to solve and create further mysteries of their own. Moreover, "extra factor" strategies have tended to reinforce the widespread criticism that libertarian notions of free will requiring indeterminism are mysterious and have no place in the modern scientific picture of the world.

So my own belief is that, if we are going to make progress on the Intelligibility Problem about libertarian free will, we must strike out in new directions, trying to avoid to the degree possible appeals to extra factor strategies, including special forms of agent-causation, and appealing to such extra factors *only* if we cannot possibly avoid them. But doing this, I believe, means rethinking issues about indeterminism and responsibility, and hence libertarian free will, from the ground up – a task to which I now turn.

## 7 Indeterminism and Responsibility

The first step in this rethinking about the Intelligibility Problem is to note that indeterminism does not have to be involved in *all* acts done "of our own free wills" for which we are ultimately responsible, as noted earlier. All free acts do not have to be undetermined on the libertarian view, but only those acts by which we made ourselves into the kinds of persons we are, namely the "will-setting" or "self-forming actions" (SFAs) that are required for ultimate responsibility.

Now I believe these undetermined self-forming actions or SFAs occur at those difficult times of life when we are torn between competing visions of what we should do or become. Perhaps we are torn between doing the moral thing or acting from ambition, or between powerful present desires and long-term goals, or we are faced with difficult tasks for which we have aversions. In all such cases, we are faced with competing motivations and have to make an effort to overcome temptation to do something else we also strongly want. There is tension and uncertainty in our minds about what to do at such times, I suggest, that is reflected in appropriate regions of our brains by movement away from thermodynamic equilibrium – in short, a kind of "stirring up of chaos" in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation is thus reflected in the indeterminacy of our neural processes themselves. What we experience internally as uncertainty about what to do on such occasions would then correspond physically to the opening of a window of opportunity that temporarily screens off complete determination by influences of the past.

When we do decide under such conditions of uncertainty, the outcome would not be determined because of the preceding indeterminacy – and yet the outcome can be willed (and hence rational and voluntary) either way owing to the fact that in such self-formation, the agents' prior wills are divided by conflicting motives. Consider a businesswoman who faces such a conflict. She is on her way to an important meeting when she observes an assault taking place in an alley. An inner struggle ensues between her conscience, to stop and call for help, and her career ambitions, which tell her she cannot miss this meeting. She has to make an effort of will to overcome the temptation to go on. If she overcomes this temptation, it will be the result of her effort, but if she fails, it will be because she did not *allow* her effort to succeed. And this is due to the fact that, while she willed to overcome temptation, she also willed to fail, for quite different and incommensurable reasons. When we, like the woman, decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we *make* one set of

competing reasons or motives prevail over the others then and there *by deciding*.

Now let us add a further piece to the puzzle. Just as indeterminism need not undermine rationality and voluntariness of choices, so indeterminism in and of itself need not undermine control and responsibility. Suppose you are trying to think through a difficult problem, say a mathematical problem, and there is some indeterminacy in your neural processes complicating the task – a kind of chaotic background. It would be like trying to concentrate and solve a problem, say a mathematical problem, with background noise or distraction. Whether you are going to succeed in solving the problem is uncertain and undetermined because of the distracting neural noise. Yet, if you concentrate and solve the problem nonetheless, we have reason to say you did it and are responsible for it, even though it was undetermined whether you would succeed. The indeterministic noise would have been an obstacle that you overcame by your effort.

There are numerous examples supporting this point, where indeterminism functions as an obstacle to success without precluding responsibility. Included among these examples are the Austin-style examples discussed in section 5. Recall the assassin who is trying to shoot the prime minister, but might miss because of some undetermined events in his nervous system that may lead to a jerking or wavering of his arm. If the assassin does succeed in hitting his target, despite the indeterminism, can he be held responsible? The answer is clearly yes because he intentionally and voluntarily succeeded in doing what he was *trying* to do – kill the prime minister. Yet his action, killing the prime minister, was undetermined. Indeterminism, it would appear, does not necessarily rule out responsibility.

Here is another example: A husband, while arguing with his wife, in a fit of rage swings his arm down on her favorite glass-top table top intending to break it. Again, we suppose that some indeterminism in his outgoing neural pathways makes the momentum of his arm indeterminate, so that it is undetermined whether the table will actually break right up to the moment when it is struck. Whether the husband breaks the table is undetermined and yet he is clearly responsible, if he does break it. (It would be a poor excuse to offer his wife, if he claimed: "Chance did it, not me." Though indeterminism was involved, chance didn't do it, he did.) In this example, as in the previous one, the agent can be held responsible for an action even though the action was undetermined.

Now these examples – of the mathematical problem, the assassin and the husband – are not all we want for free will, since they do not amount to genuine exercises of self-forming actions (SFAs), like the businesswoman's, where the will is divided between conflicting motives. The businesswoman wants to help the victim, but she also wants to go on to her meeting. By

contrast, the assassin's will is not equally divided. He wants to kill the prime minister, but he does not also want to fail. (If he fails therefore, it will be *merely* by chance.) Yet these examples of the assassin, the husband and the like, while they do not tell us all we want to know about free will, do provide some clues about what free will requires. To go further, we have to add some additional twists.

## 8 Parallel Processing

Imagine in cases of conflict characteristic of self-forming actions or SFAs, like the businesswoman's, that the indeterministic noise which is providing an obstacle to her overcoming temptation is not coming from an external source, but has its source in her own will, since she also deeply desires to do the opposite. To understand how this could be, imagine that two crossing recurrent neural networks are involved in the brain, each influencing the other, and representing her conflicting motivations. (Recurrent neural networks are complex networks of interconnected neurons in the brain circulating impulses in feedback loops that are generally involved in higher-level cognitive processing.) The input of one of these neural networks consists in the woman's reasons for acting morally and stopping to help the victim; the input of the other network comprises her ambitious motives for going on to her meeting.

The two networks are connected so that the indeterminism that is an obstacle to her making one of the choices is present because of her simultaneous conflicting desire to make the other choice – the indeterminism thus arising from a tension-creating conflict in the will, as we said. This conflict, as noted earlier, would be reflected in appropriate regions of the brain by movement away from thermodynamic equilibrium. The result would be a stirring up of chaos in the neural networks involved. Chaos in physical systems is a phenomenon in which very small changes in initial conditions are magnified so that they lead to large and unpredictable changes in the subsequent behavior of a system. You may have heard the popular illustration of chaos in which the fluttering of a butterfly's wings in South America initiates a chain of events that affects the weather patterns of North America. Such popular examples may be an exaggeration. But chaotic phenomena, in which small changes lead to large effects, are now known to be far more common in nature than previously believed; and they are particularly common in living things. There is growing evidence that chaos plays a role in the information processing of the brain, providing some of the flexibility that the nervous system needs to adapt creatively – rather than in predictable or rigid ways – to an ever-changing environment.

Now determinists are quick to point out that chaos, or chaotic behavior, in physical systems, though *unpredictable*, is usually deterministic and does not itself imply genuine indeterminism in nature. But some scientists have suggested that a combination of chaos and quantum physics might provide the genuine indeterminism one needs. If the processing of the brain does "make chaos in order to make sense of the world" (as one recent research paper puts it), then the resulting chaos might magnify quantum indeterminacies in the firings of individual neurons so that they would have large-scale indeterministic effects on the activity of neural networks in the brain as a whole. If chaotic behavior were thus enhanced in these neural networks by tension-creating conflict in the will, the result would be some significant indeterminism in the cognitive processing of each of the competing neural networks.

In such circumstances, when either of the competing networks "wins" (or reaches an activation threshold, which amounts to choice), it would be like your solving the mathematical problem by overcoming the background indeterministic noise created by the presence of the competing network. And just as when you solved the mathematical problem by overcoming the distracting noise, one can say you did it and are responsible for it, so one can also say this, I would argue, in the present case, *whichever outcome is chosen*. For the neural pathway through which the woman does succeed in reaching a choice threshold will have overcome the obstacle in the form of indeterministic noise generated by the presence of the other competing network.

Note that, under such conditions, the choice the woman might make either way will not be "inadvertent," "accidental," "capricious," or "merely random" (as critics of indeterminism say) because the choice will be *willed* by the woman either way when it is made, and it will be done for *reasons* either way – reasons that she then and there *endorses*. For, let us recall that in SFAs, the agent's will is divided and the agent has strong reasons or motives for making *either* choice. So when she decides, she endorses one set of competing reasons over the other as the one she will act on. But *willing* what you do in this way, and doing it for *reasons* that you endorse, are conditions usually required to say something is done "on purpose," rather than accidentally, capriciously, or merely by chance. Moreover, these conditions taken together (that the choices were willed either way, were done for reasons and the agents endorsed them) rule out each of the reasons we have for saying that agents act, but do not have *control* over their actions. The businesswoman's choice either way, for example, will not have been made accidentally or inadvertently or by mistake, nor need it have been the result of coercion (no one was holding a gun to her head, for example) or the result of control by other agents. Of course, for undetermined SFAs, agents do not control or determine which choice outcome will occur *before* it occurs. But it does not follow, because one does not control

or determine which of a set of outcomes is going to occur before it occurs, that one does not control or determine which of them occurs, *when* it occurs. When the above conditions for SFAs are satisfied, agents exercise control over their future lives *then and there* by deciding.

As a consequence, they have what I call *plural voluntary control* over their options in the following sense: Agents have plural voluntary control over a set of options (such as the woman's choosing to help the victim or to go on to her meeting), when they are able to bring about *whichever* of the options they will, *when* they will to do so, *for* the reasons they will to do so, *on* purpose, rather than accidentally or by mistake, *without* being coerced or compelled in doing so or willing to do so, or otherwise controlled in doing or willing to do so by any other agents or mechanisms. Each of these conditions can be satisfied for SFAs, like the businesswoman's, as I have described them. The conditions can be summed up by saying that the agents can choose either way *at will*. In other words, the choices are "will-setting": We set our wills one way or the other in the *act* of deciding itself, and not before.

Note also that this account of self-forming choices or SFAs amounts to a kind of "doubling" of the mathematical problem. It is as if an agent faced with such a self-forming choice is *trying* or making an effort to solve *two* cognitive problems at once, or to complete two competing (deliberative) tasks at once – in our example, to make a moral choice and to make a conflicting self-interested choice (corresponding to the two competing neural networks involved). Each task is being thwarted by the indeterminism generated by the presence of the competing network, so it might fail. But if it succeeds, then the agents can be held responsible because, as in the case of solving the mathematical problem, the agents will have succeeded in doing what they were knowingly and willingly trying to do.

Recall the assassin and the husband. Owing to indeterminacies in their neural pathways, the assassin might miss his target or the husband fail to break the table. But if they *succeed*, despite the probability of failure, they are responsible, because they will have succeeded in doing what they were trying to do. And so it is, I suggest, with self-forming choices (SFAs) like the businesswoman's, except that in the case of self-forming choices, *whichever way the agents choose* they will have succeeded in doing what they were trying to do because they were simultaneously trying to make both choices, and one is going to succeed. Their failure to do one thing is not a *mere* failure, but a voluntary succeeding in doing the other.

Does it make sense to talk about the agent's trying to do two competing things at once in this way, or to solve two cognitive problems at once? Well, we now know that the brain is a "parallel processor"; it can simultaneously process different kinds of information relevant to tasks such as perception or recognition through different neural pathways. Such a capacity, I believe, is

essential to the exercise of free will. In cases of self-formation (SFAs), agents are simultaneously trying to resolve plural and competing cognitive tasks. They are, as we say, of two minds. Yet they are not two separate persons. They are not dissociated from either task. The businesswoman who wants to go back to help the victim is the same ambitious woman who wants to go to her meeting and make a sale. She is torn inside by different visions of who she is and what she wants to be, as we all are from time to time. But this is the kind of complexity needed for genuine self-formation and free will. And when she succeeds in doing one of the things she is trying to do, she will endorse that outcome as *her* resolution of the conflict in her will, voluntarily and intentionally, not by accident or mistake.

## 9 Responsibility, Luck, and Chance

You may find all this interesting and yet still find it hard to shake the intuition that if choices are undetermined, they *must* happen merely by chance – and so must be "random," "capricious," "uncontrolled," "irrational," and all the other things usually charged. Such intuitions are deeply ingrained. But if we are going to understand free will, I think we must break old habits of thought supporting such intuitions and learn to think in new ways.

The first step is to question the intuitive connection in people's minds between "indeterminism's being involved in something" and "its happening merely as a matter of chance or luck." "Chance" and "luck" are terms of ordinary language that carry the meaning of "its being out of my control." So using them already begs certain questions. Whereas "indeterminism" is a technical term that merely rules out *deterministic* causation, though not causation altogether. Indeterminism is consistent with nondeterministic or probabilistic causation, where the outcome is not inevitable. It is therefore a mistake (in fact, one of the most common in debates about free will) to assume that "undetermined" means "uncaused" or "*merely* a matter of chance."

Here is another source of misunderstanding. Since the outcome of the businesswoman's effort (the choice) is undetermined up to the last minute, we may have the image of her first making an effort to overcome the temptation to go on to her meeting and then at the last instant "chance takes over" and decides the issue for her. But this is a mistaken image. On the view just presented, one cannot separate the indeterminism and the effort of will, so that *first* the effort occurs *followed* by chance or luck (or vice versa). One must think of the effort and the indeterminism as fused; the effort *is* indeterminate and the indeterminism is a property of the effort, not something separate that occurs after or before the effort. The fact that the effort has this property of being indeterminate does not make it any less the woman's *effort*. The complex

recurrent neural network that realizes the effort in the brain is circulating impulses in feedback loops and there is some indeterminacy in these circulating impulses. But the whole process is her effort of will and it persists right up to the moment when the choice is made. There is no point at which the effort stops and chance "takes over." She chooses as a result of the effort, even though she might have failed. Similarly, the husband breaks the table as a result of his effort, even though he might have failed because of the indeterminacy. (That is why his excuse, "chance broke the table, not me," is so lame.)

Just as expressions like "she chose *by* chance" can mislead us in such contexts, so can expressions like "she got lucky." Recall that one might say of the assassin and husband "they got lucky" in killing the prime minister and breaking the table because their actions were undetermined. Yet the surprising thing is that we can still say the assassin and husband were *responsible* if they succeeded in killing the prime minister and breaking the table. So we should ask ourselves the following question: why is it wrong to say "he got lucky, *so he was not responsible*" in the cases of the husband and the assassin? For it *is* wrong to say this since they did get lucky and yet they were *still* responsible. (Imagine the assassin's lawyer arguing in the courtroom that his client is not guilty because his killing the prime minister was undetermined and might therefore have failed by chance. Would such a defense succeed?)

The first part of an answer as to why the assassin and husband are still responsible has to do with the point made earlier about "luck" and "chance." These two words have question-begging implications in ordinary language that are not necessarily implications of "indeterminism" (which implies only the absence of deterministic causation). The core meaning of "he got lucky" in the assassin and husband cases, which *is* implied by indeterminism, is that "he succeeded *despite the probability or chance of failure*"; and this core meaning does not imply lack of responsibility, *if he succeeds*. If "he got lucky" had other meanings in these cases that are often associated with "luck" and "chance" in ordinary usage, the inference "he got lucky so he was not responsible" would not fail for the husband and assassin, as it clearly does. For example, if "luck" in these cases meant the outcome was not his doing, or occurred by mere chance, or he was not responsible, then the inference "he got lucky so he was not responsible" would hold for the husband and assassin. But the point is that these further meanings of "luck" and "chance" do not follow *from the mere presence of indeterminism*.

The second reason why the inference "he got lucky, so he was not responsible" does not work in the cases of the assassin and the husband is that *what* they succeeded in doing was what they were *trying* and wanting to do all along (kill the minister and break the table respectively). The third reason is that *when* they succeeded, their reaction was not "Oh dear, that was a mistake,

an accident – something that *happened* to me, not something I *did*." Rather they *endorsed* the outcomes as something they were trying and wanting to do all along, knowingly and purposefully, not by mistake or accident.

But these conditions are satisfied in the businesswoman's case as well *either way* she chooses. If she succeeds in choosing to return to help the victim (or in choosing to go on to her meeting) (i) she will have "succeeded *despite the probability or chance of failure*," (ii) she will have succeeded in doing what she was *trying* and *wanting* to do all along (she wanted both outcomes very much, but for different reasons, and was trying to make those reasons prevail in both cases), and (iii) when she succeeded (in choosing to return to help) her reaction was not "Oh dear, that was a mistake, an accident – something that happened to me, not something I did." Rather she *endorsed* the outcome as something she was trying and wanting to do all along; she recognized the choice as her resolution of the conflict in her will. And if she had chosen to go on to her meeting she would have endorsed that outcome, recognizing it as her resolution of the conflict in her will.

## 10 Choice, Agency, Efforts, and Causes: Further Objections Considered

Perhaps we are begging the question by assuming the outcomes of the woman's efforts are *choices* to begin with. If indeterminism is involved in a process (such as the woman's deliberation) so that its outcome is undetermined, one might argue that the outcome must merely *happen* and therefore cannot be somebody's *choice*. But there is no reason to assume such a claim is true. A choice is the formation of an intention or purpose to do something. It resolves uncertainty and indecision in the mind about what to do. Nothing in such a description implies that there could not be some indeterminism in the deliberation and neural processes of an agent preceding choice corresponding to the agent's prior uncertainty about what to do. Recall from the preceding arguments that the presence of indeterminism does not mean the outcome happened *merely* by chance and *not* by the agent's effort. Self-forming choices are undetermined, but not uncaused. They are caused by the agent's efforts.

Well, perhaps indeterminism does not undermine the idea that something is a *choice* simply, but rather that it is the *agent's* choice. This objection raises important questions about agency. What makes the woman's choice her own on the above account is that it results from *her* efforts and deliberation, which in turn are causally influenced by her reasons and her intentions (for example, her intention to resolve indecision in one way or another). And what makes these efforts, deliberation, reasons, and intentions *hers* is that they are embedded in a larger motivational system realized in her brain in terms of which

she defines herself as a practical reasoner and actor. A choice is the agent's when it is produced intentionally by efforts, by deliberation and by reasons that are part of this self-defining motivational system and when, in addition, the agent *endorses* the new intention or purpose created by the choice into that motivational system as a further purpose to guide *future* practical reasoning and action.

Another concern that has been raised about the above account of libertarian free will is that we are not introspectively aware of making dual efforts and performing multiple cognitive tasks in such choice situations. But I am not claiming that agents are conscious of making dual efforts. What they are introspectively conscious of is that they are trying to decide about which of two options to choose and that either choice is a difficult one because there are resistant motives pulling them in different directions that will have to be overcome, whichever choice is made. In such introspective conditions, I am theorizing that what is actually going on underneath is a kind of parallel processing in the brain that involves separate efforts or endeavorings to resolve competing cognitive tasks. The point is that introspective evidence does not give us the whole story about free will. If we stay on the surface and just consider what our immediate experience tells us, free will, I believe, is bound to appear mysterious, as it has appeared to so many people through the centuries. To unravel its mysteries, we have to consider what might be going on behind the scenes.

It is now widely believed, for example, that parallel processing takes place in the brain in such cognitive phenomena as visual perception. The theory is that the brain separately processes different features of the visual scene, such as object and background, through distinguishable and parallel, though interacting, neural pathways or streams. Suppose someone objected that we are not introspectively aware of such distributed processing in ordinary cases of perception. That would hardly be a decisive objection to this new theory of vision. For the claim is that this is what we are doing in visual perception, not necessarily that we are introspectively aware of doing it. And I am making a similar claim about free will. What is needed is a *theory* about what might be going on when we exercise free will, not merely a description of what we immediately experience.

It has also been objected that it is irrational to make efforts to do incompatible things. I concede that in most ordinary situations it is. But building on suggestions made by theorists of action, such as Michael Bratman, I argue that there are special circumstances in which it is not irrational to make competing efforts: These include circumstances in which: (i) we are deliberating between competing options; (ii) we intend to choose one or the other, but cannot choose both; (iii) we have powerful motives for wanting to choose each of the options for different and incommensurable reasons; (iv) there is

a consequent resistance in our will to either choice, so that (v) if either choice is to have a chance of being made, effort will have to be made to overcome the temptation to make the other choice; and, most importantly, (vi) we want to give each choice a fighting chance of being made because the motives for each choice are important to us. The motives for each choice define in part what sort of person we are; and we would taking them lightly if we did not make an effort in their behalf. These conditions are, of course, the conditions of SFAs.

But perhaps the deepest concern about the above theory remains the concern about *chance*. If chance is involved in decision making, we somehow think of chance as deciding the issue, like spinning a wheel to select an outcome. As noted earlier, that worry sends us scurrying around looking for extra factors, other than prior events or happenings, to tip the balance to one choice or the other, such as an immaterial agent or non-event agent cause. But there is an alternative way to think about the way that indeterminism might be involved in free choice that first occurred to me twenty-five years ago, a way that avoids these familiar libertarian stratagems and requires a transformation of perspective.

Think, instead, of the indeterminism involved in free choice as an *ingredient* in a larger goal-directed or teleological process or activity, in which the indeterminism functions as a *hindrance* or *obstacle* to the attainment of the goal. If you reflect for a moment, you will see that this is what the account of free will presented earlier is actually doing. Here is an example from another modern scientific theory of relevance to free will – namely, information theory. Consider the sending of a message in Morse code. The sender taps out the message in dots and dashes, representing letters. The pulses travel electrically over lines to the receiver where they are reproduced. Now, there may be interference due to noise or static in the electrical lines so that the message does not get through, or a distorted message gets through. In that case we have what information theorists call "equivocation" rather than mere noise. The message is too garbled to read. If the message does get through, however, despite the electrical noise or static, then the goal of the message sender is realized. Now if the noise in the electrical lines were the result of indeterminism or chance, whether the message gets through would be undetermined. Yet if the undetermined electrical noise or static was not great enough to cause equivocation, the goal of the process would be realized, despite the interference (the message would get through despite the indeterminism).

In a similar fashion, the idea is not to think of the indeterminism involved in free choices as a cause *acting on its own*, but as an ingredient in a larger goal-directed or teleological process or activity in which the indeterminism functions as a hindrance or obstacle to the attainment of the

goal. This is the role suggested for indeterminism in the efforts preceding undetermined SFAs. These efforts are temporally extended goal-directed activities in which indeterminism is a hindering or interfering element, like the noise or static in the message transmission example. The choices, or SFAs, that result from these temporally extended activities or efforts thus do not pop up out of nowhere, even though they are undetermined. They are the *achievements* of goal-directed activities of the agent that might have failed, but did not.

Note that, if indeterminism or chance does play this kind of interfering role in a larger process leading to choice, the indeterminism or chance need not be the *cause* of the choice that is actually made. This follows from a general point about probabilistic causation. A vaccination may hinder or lower the probability that I will get a certain disease, so it is causally relevant to the outcome. But if I get the disease despite it, the vaccination is not the *cause* of my getting the disease, though it was causally relevant, because its role was to *hinder* that effect. The causes of my getting the disease, by contrast, are those causally relevant factors (such as the infecting virus) that significantly *raised* the probability of its occurrence. Similarly, in the case of the businesswoman's choice, the causes of the choice she does make (the moral choice or the ambitious choice) are those causally relevant factors that significantly *raised* the probability of making *that* choice from what it would have been if those factors had not been present, such as her reasons and motives for making that choice rather than the other, her conscious awareness of these reasons and her deliberative efforts to overcome the temptations to make the contrary choice. The presence of the indeterminism lowers the probability that the choice will result from these reasons, motives, and efforts from what that probability would have been if there had been no competing motives or efforts and hence no interfering indeterminism.

Since those causally relevant features of the agent, which *can* be counted among the causes of the woman's choice, are *her* reasons or motives, *her* conscious awareness and *her* deliberative efforts, we can also say that she is the cause of the choice by virtue of making the efforts for the reasons and succeeding. The indeterminism or chance (like the vaccination) was causally relevant to the outcome, but it was not the cause. This explains why the husband's excuse was so lame when he said "Chance broke the table, not me." While chance was causally involved, chance was not the cause of the table's breaking. The cause was his effort to break the table by swinging his arm down on it. The chance merely made it uncertain whether that larger goal-directed activity would succeed. And so it is, I suggest, with the efforts leading to self-forming choices. These efforts, of course, are mental activities realized in the higher cognitive processing of the brain rather than in overt actions such as the swinging of an arm. But the SFAs that result from

these mental efforts are nonetheless also the achievements of goal-directed activities that might have failed due to chance, but did not, just as the husband's effort to break the table by swinging his arm might have failed due to chance, but did not.

But can't we say that it is a "matter of chance" whether one of these efforts leading to SFAs succeeds or not? For isn't it true that whether or not an effort succeeds in producing a choice depends on whether certain undetermined neurons involved in the agent's cognitive processing fire or do not fire (perhaps within a given time frame)? And whether these neurons fire or not is by hypothesis undetermined, is it not, and therefore not under the control of the agent? Well, yes, we *can* say all of these things: whether an effort succeeds *does* depend upon whether certain undetermined neurons fire or not; and whether these neurons fire is not under the control of the agent; and we can consequently say it is a matter of chance whether the efforts leading to SFAs succeed or not.

But the really astonishing thing is that, while all these things can be truly said, it *does not follow* that the agent is not *responsible* for the choice, *if* the effort succeeds. For, consider the husband swinging his arm down on the table. It is *also* true in his case that whether or not his effort to break the table succeeds "depends" on whether certain neurons in his arm fire or do not fire; and it is *also* true in his case that whether these neurons fire or not is undetermined and therefore not under his control; and we can *also* consequently say in the husband's case that it is a "matter of chance" whether or not he succeeds in breaking the table. Yet, even though we can say all this, it does not follow that he is not responsible for breaking the table, *if* his effort succeeds. Astonishing indeed! But this is the kind of surprising result one gets when indeterminism or chance plays an interfering or hindering role in larger goal-directed activities, such as efforts to do certain things that may succeed or fail.

It is well to meditate on this: We tend to reason that if an outcome (breaking a table *or* making a choice) depends on whether certain neurons fire or not (in the arm *or* in the brain), then the agent must be able to *make* those neurons fire or not, if the agent is to be responsible for the outcome. In other words, we think we have to crawl down to the place where the indeterminism originates (in the individual neurons) and *make* them go one way or the other. We think we have to become originators at the micro-level and tip the balance that chance leaves untipped, if we (and not chance) are to be responsible for the outcome. And we realize, of course, that we can't do that. But we don't have to. It's the wrong place to look. We don't have to micro-manage our individual neurons one by one to perform purposive actions. In fact, we do not have such micro-control over our neurons *even when we perform ordinary actions* such as swinging an arm down on a table.

What we need when we perform purposive activities, mental or physical, is rather macro-control of processes involving many neurons – complex processes that may succeed in achieving their goals despite the interfering effects of some recalcitrant neurons. We don't micro-manage our actions by controlling each individual neuron or muscle that might be involved. We don't know enough about neurology or physiology to do that; and it would be counterproductive to try. But that does not prevent us from macro-managing our purposive activities (whether they be mental activities such as practical reasoning, or physical activities, such as arm-swingings) and being responsible when those purposive activities attain their goals.

## 11 Responsibility and Control: Three Assassins

But does not the presence of indeterminism or chance at least *diminish* the control persons have over their choices or actions? And would that not affect their responsibility? (This is another way in which objections about chance and luck have often been raised against libertarian views of free will.) Is it not the case that the assassin's control over whether the prime minister is killed (his ability to realize his purposes or what he is trying to do) is lessened by the undetermined impulses in his arm – and so also for the husband and his breaking the table? The answer is yes, again. But the further surprising point worth noting – a point that I think is so often missed – is that *diminished control* in such circumstances *does not entail diminished responsibility* when the agents succeed in doing what they are trying to do. Ask yourself this question: Is the assassin less guilty of killing the prime minister, if he did not have complete control over whether he would succeed because of the indeterminism in his neural processes?

Suppose there were three assassins, each of whom killed a prime minister. Suppose one of them had a 50 percent chance of succeeding because of the indeterministic wavering of his arm. Another had an 80 percent chance, and the third a 100 percent chance. (With this third assassin there was no wavering at all; he was a young stud assassin.) Is one of these assassins less guilty than the other, *if they all succeed*? Should we say that one assassin deserves a hundred years in jail, the other eighty years and the third fifty years? Absurd. They are all equally guilty if they succeed. The diminished control in the assassins who had an 80 percent or a 50 percent chance does not translate into diminished responsibility when they succeed. Diminished control in such circumstances does not entail diminished responsibility. Imagine a lawyer for the 50 percent assassin arguing that his client was not guilty because the prime minister's dying as a result of what his client did was a "matter of chance." Therefore chance was the cause of the prime minister's death, not

his client. That would make the notorious "Twinkie Defense" look brilliant by comparison. (This was the defense offered by a lawyer in California that his client was not responsible because the client's blood sugar was so high from having eaten too many Twinkies that he could not control his actions.)

There is an important further lesson here I believe about free will in general. We should concede that indeterminism, wherever it occurs, *does* diminish control over what we are trying to do and *is* a hindrance or obstacle to the realization of our purposes. But recall that in the case of the businesswoman (and SFAs generally), the indeterminism that is admittedly diminishing her control over one thing she is trying to do (the moral act of helping the victim) *is coming from her own will* – from her desire and effort to do the opposite (go to her business meeting). And the indeterminism that is diminishing her control over the other thing she is trying to do (act selfishly and go to her meeting) is coming from her desire and effort to do the opposite (to be a moral person and act on moral reasons). In each case, the indeterminism *is* functioning as a hindrance or obstacle to her realizing one of her purposes – a hindrance or obstacle in the form of resistance within her will which has to be overcome by effort.

If there were no such hindrance – if there were no resistance in her will – she would indeed in a sense have "complete control" over one of her options. There would be no competing motives standing in the way of her choosing it and therefore no interfering indeterminism. But then also, she would not be free to rationally and voluntarily choose the *other* purpose because she would have no good competing reasons to do so. Thus, by *being* a hindrance to the realization of some of our purposes, indeterminism paradoxically opens up the genuine possibility of pursuing other purposes – of choosing or doing *otherwise* in accordance with, rather than against, our wills (voluntarily) and reasons (rationally). To be genuinely self-forming agents (creators of ourselves) – to have free will – there must at times in life be obstacles and hindrances in our wills of this sort that we must overcome.

I think libertarians about free will have traditionally tried to ignore this aspect of indeterminism. They knew indeterminism was required on their view, but assumed it could be entirely circumvented by special agencies. But hindrances and obstacles and resistance in the will are precisely what are needed for free will, which, like life itself, exists near the edge of chaos. If one were to put it in a religious perspective, this fact would be related to the problem of evil. There must be hindrances and obstacles to our choices and resistance in our own wills to be overcome, if we are to be capable of genuine self-formation and free will. Compare Evodius's question to St Augustine (in Augustine's classic work *On the Free Choice of the Will*) of why God gave us free will since it brings so much conflict, struggle and suffering into the world.

Yes, it does bring struggle, hindrances and resistance in our wills. But such things are necessary for genuine responsibility.

Of interest also is Kant's image, which I have used before, of the bird that is upset by the resistance of the air and the wind to its flight and so imagines that it could fly better if there were no air at all to resist it. But of course the bird would not fly better if there were no air. It would cease to fly at all. So it is with indeterminism in relation to free will. It provides resistance to our choices, but a resistance that is necessary if we are to be capable of true self-formation.

## 12 Conclusion: Complexity and "Being an Author of One's Own Story"

In summary, I think the key to understanding the role of chance in free will is not to think of chance as a causal factor by itself, but rather to think of chance as an interfering ingredient in larger goal-directed processes. Viewing chance in this way is related to a peculiarly modern scientific way of understanding human agency that also his its roots in the ancient view of Aristotle. Agents, according to this modern conception with ancient roots, are to be conceived as *information-responsive complex dynamical systems*. Complex dynamical systems are the subject of "dynamical systems theory" and also of what is sometimes popularly called "complexity theory." They are systems (which are now known to be ubiquitous in nature) in which new *emergent* capacities arise as a result of greater complexity or as the result of movement away from thermodynamic equilibrium toward the edge of chaos. When these emergent capacities arise in complex dynamical systems, the systems as a whole impose novel constraints on the behavior of their parts that did not constrain the parts before the new complexity or disequilibrium was achieved. In such complex dynamical systems there is thus a reciprocal causal influence of wholes to parts and parts to wholes.

In the account of free will I have proposed, for example, it is a conflict in the larger motivational system of the agent taken as a whole – the self-network, as I have elsewhere called it – that stirs up chaos and amplifies indeterminism at the neuronal and synaptic levels. The larger whole or self-network thus stirs up chaos in its parts (neurons and networks of neurons), but the resulting amplified indeterminism in turn interferes with the goal-directed activities of the larger network. There is thus a mutual influence of wholes to parts and parts to wholes characteristic of complex dynamical systems. And emergent capacities are also involved. Only when creatures attain the kind of inner complexity capable of giving rise to conflicts in their wills, or motivational systems, between incommensurable values does

the capacity for self-formation characteristic of free will arise. So we are talking about a special kind of complex dynamical system that is information-responsive in highly complex ways, not seen in non-rational animals. The businesswoman, as I said, is torn inside by different visions of who she is and what she wants to be, as we all are from time to time. But this is just the kind of complexity needed for the novel capacity of genuine self-formation or free will to emerge.

Let me conclude with one final objection to the account of free will presented here, which is perhaps the most telling and has not yet been discussed. Even if one granted that persons, such as the businesswoman, could make genuine self-forming choices that were undetermined, isn't there something to the charge that such choices would be *arbitrary*? A residual arbitrariness seems to remain in all self-forming choices since the agents cannot in principle have sufficient or conclusive *prior* reasons for making one option and one set of reasons prevail over the other.

There is some truth to this objection also, but again I think it is a truth that tells us something important about free will. It tells us that every undetermined self-forming free choice is the initiation of what might be called a *value experiment* whose justification lies in the future and is not fully explained by past reasons. In making such a choice we say, in effect, "Let's try this. It is not required by my past, but it is consistent with my past and is one branching pathway in the garden of forking paths my life can now meaningfully take. Whether it is the right choice, only time will tell. Meanwhile, I am willing to take responsibility for it one way or the other."

It is worth noting that the term "arbitrary" comes from the Latin *arbitrium*, which means "judgment" – as in *liberum arbitrium voluntatis*, "free judgment of the will" (the medieval philosophers' designation for free will). Imagine a writer in the middle of a novel. The novel's heroine faces a crisis and the writer has not yet developed her character in sufficient detail to say exactly how she will act. The author makes a "judgment" about this that is not determined by the heroine's already formed past which does not give unique direction. In this sense, the judgment (*arbitrium*) of how she will react is "arbitrary," but not entirely so. It had input from the heroine's fictional past and in turn gave input to her projected future. In a similar way, agents who exercise free will are both authors of and characters in their own stories all at once. By virtue of "self-forming" judgments of the will (*arbitria voluntatis*) (SFAs), they are "arbiters" of their own lives, "making themselves" out of past that, if they are truly free, does not limit their future pathways to one.

Suppose we were to say to such persons: "But look, you didn't have sufficient or *conclusive* prior reasons for choosing as you did since you also had viable reasons for choosing the other way." They might reply. "True enough. But I did have *good* reasons for choosing as I did, which I'm willing to stand

by *and take responsibility for*. If these reasons were not sufficient or conclusive reasons, that's because, like the heroine of the novel, I was not a fully formed person before I chose (and still am not, for that matter). Like the author of the novel, I am in the process of writing an unfinished story and forming an unfinished character who, in my case, is myself."

## Further Reading

For more advanced discussion of the issues discussed in this chapter, see the collection of readings in Robert Kane (ed.), *The Oxford Handbook of Free Will* (Oxford: Oxford University Press, 2002). The following collections of essays also contain further readings on the issues about free will discussed in this chapter: Gary Watson (ed.), *Free Will* (Oxford: Oxford University Press, 2003), Robert Kane (ed.), *Free Will* (Oxford: Blackwell Publishers, 2002), and Laura Ekstrom (ed.), *Agency and Responsibility: Essays on the Metaphysics of Freedom* (Boulder, CO: Westview Press, 2001). Another collection of readings which deals specifically with libertarian accounts of free will is Timothy O'Connor (ed.), *Agents, Causes and Events: Essays on Free Will and Indeterminism* (Oxford: Oxford University Press, 1995).

The libertarian view of free will presented in this chapter is further developed in my book *The Significance of Free Will* (Oxford: Oxford University Press, 1996). Alternative accounts of libertarian free will may also be found in Timothy O'Connor, *Persons and Causes* (Oxford: Oxford University Press, 2000); Randolph Clarke, *Libertarian Accounts of Free Will* (Oxford: Oxford University Press, 2003); Carl Ginet, *On Action* (Cambridge: Cambridge University Press, 1990); Hugh McCann, *The Works of Agency* (Ithaca, NY: Cornell University Press, 1998), Stewart Goetz, "A Non-causal Theory of Agency," *Philosophy and Phenomenological Research* 49 (1988), 303–16; Laura Waddell Ekstrom, *Free Will* (Boulder, CO: Westview Press, 2000); David Hodgson "A Plain Person's Free Will," *Journal of Consciousness Studies* 12 (2005), 3–19; James Felt, *Making Sense of Our Freedom* (Ithaca, NY: Cornell University Press, 1994), and Thomas Pink, *Free Will: A Short Introduction* (Oxford: Oxford University Press, 2004). O'Connor and Clarke defend sophisticated modern versions of the "agent-causation theory" of libertarian free will, which was mentioned in the chapter. Ginet, McCann, and Goetz defend what are called "simple indeterminist" or "noncausalist" accounts of libertarian free will. The view defended in this chapter is often called a "causal indeterminist" or "event-causal" view of libertarian free will to distinguish it from "agent-causation" and "simple indeterminist" views. Ekstrom also defends a causal indeterminist view, though different than the one defended in this chapter. The libertarian views defended by Hodgson, Felt, and Pink are not easily fitted into any of these three familiar categories.

Readable introductions for the non-specialist about the role of *neural networks* (including *recurrent neural networks*) in cognitive processing and neuroscience include P. M. Churchland, *The Engine of Reason, the Seat of the Soul* (Cambridge, MA: MIT Press, 1996) and Manfred Spitzer, *The Mind Within the Net* (Cambridge, MA: MIT Press, 1999). On the role of *chaos* and chaotic processes in the brain, see, e.g. C. Skarda and W. Freeman, "How Do Brains Make Chaos in Order to Make Sense of the World?" *Behavioral and Brain Sciences* (1987) 10: 161–95 and H. Walter, *Neurophilosophy of Free Will* (Cambridge, MA: MIT Press, 2001, Part III). An overview of research on *parallel processing* in visual perception can be found in "Decomposing and Localizing Vision: An Exemplar for Cognitive Neuroscience" by William Bechtel, in *Philosophy and the Neurosciences: A Reader*, ed. by W. Bechtel, Pete Mandik, Jennifer Mundale and Robert Stufflebaum (Oxford: Blackwell Publishers, 2001), pp. 225–49. Introductions to *complex dynamical systems* or *complexity* for non-specialists include R. Lewin, *Complexity: Life at the Edge of Chaos* (New York: Macmillan Publishers, 1992) and M. Mitchell Waldrop, *Complexity: The Emerging Science at the Edge of Chaos* (New York: Simon and Schuster, 1992). Works that attempt to apply theories about complex dynamical systems to issues about action and agency include E. Thelen and R. B. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action* (Cambridge, MA: MIT Press, 1994) and Alicia Juarrero, *Dynamics in Action: Intentional Behavior as a Complex System* (Cambridge, MA: MIT Press, 1999).

# 2
# *Compatibilism*

*John Martin Fischer*

*Because you're mine, I walk the line . . .*
*Johnny Cash*

## 1 Introduction

Thinking about it in one way, compatibilism seems very plausible. For now, take "compatibilism" to be the doctrine that both some central notion of freedom and also genuine, robust moral responsibility are compatible with the doctrine of causal determinism (which, among other things, entails that every bit of human behavior is causally necessitated by events in the past together with the natural laws). Of course, compatibilism, as thus understood, does not in itself take any stand on whether causal determinism is true.

Compatibilism can seem plausible because it appears so obvious to us that we (most of us) are at least sometimes free and morally responsible, and yet we also realize that causal determinism could turn out to be true. That is, for all we know, it is true that all events (including human behavior) are the results of chains of necessitating causes that can be traced indefinitely into the past. Put slightly differently, I could certainly imagine waking up some morning to the newspaper headline, "Causal Determinism Is True!" (Most likely this would not be in the *National Inquirer* or even *People* – but perhaps the *New York Times* . . .) I could imagine reading the article and subsequently (presumably over some time) becoming convinced that causal determinism is true – that the generalizations that describe the relationships between complexes of past events and laws of nature, on the one hand, and subsequent events, on the other, are universal generalizations with 100 percent probabilities associated with them. And I feel confident that this would not, nor should it, change my view of myself and others as (sometimes) free and robustly morally responsible agents – deeply different from other animals. The mere

fact that these generalizations or conditionals have 100 percent probabilities associated with them, rather than 99.9 percent (say), would not and should not have any effect on my views about the existence of freedom and moral responsibility. My basic views of myself and others as free and responsible are and should be resilient with respect to such a discovery about the arcane and "close" facts pertaining to the generalizations of physics. (This of course is not to say that these basic views are resilient to *any* empirical discovery – just to this sort of discovery.)

So, when I deliberate I often take it that I am free in the sense that I have more than one option that is genuinely open to me. Since causal determinism (in the sense sketched above) might, for all we know be true, compatibilism seems extremely attractive. Similarly, it is very natural to distinguish those agents who are compelled to behave as they do from those who act freely; we make this distinction, and mark the two classes of individuals, in common sense and also the law. But since causal determinism might, for all we know, be true, compatibilism is extremely attractive. If casual determinism turned out to be true, and incompatibilism were also true, then it would seem that all behavior would be put into one class – the distinctions we naturally and intuitively draw in common sense and law would be in jeopardy of disappearing.

And yet there are deep problems with compatibilism. Perhaps these are what have led some philosophers to condemn it in such vigorous terms: "wretched subterfuge," (Kant), "quagmire of evasion" (James), and "the most flabbergasting instance of the fallacy of changing the subject to be encountered anywhere in the complete history of sophistry . . . [a ploy that] was intended to take in the vulgar, but which has beguiled the learned in our time" (Wallace Matson).

In this essay I will start by highlighting the attractions of compatibilism and sketching and motivating an appealing version of traditional compatibilism. I shall then present a basic challenge to such a compatibilism. Given this challenge, I suggest an alternative version of compatibilism, which I call "semicompatibilism," and I develop some of the advantages of such an approach. Finally, I consider objections to this specific version of compatibilism, as well as compatibilism in general. My goal will be to present the scaffolding of a defense of semicompatibilism (highlighting the main attractions), rather than a detailed elaboration or defense of the doctrine.

## 2 The Lure of Compatibilism

As I pointed out above, often it seems to me that I have more than one path open to me. The paths into the future branch out from the present, and they

represent different ways I could proceed into the future. When I deliberate now about whether to go to the lecture or to the movies tonight, I now think I genuinely can go to the lecture, and I genuinely can go to the movies (but perhaps I cannot do both). And I often have this view about the future – the view of the future as a "garden of forking paths" (in Borges' wonderful phrase). But I can also be brought to recognize that, for all I know, causal determinism is true; its truth would not necessarily manifest itself to me phenomenologically. Thus, compatibilism is extremely attractive: it allows me to keep both the view that I (sometimes at least) have more than one path genuinely open to me and also that causal determinism may be true. (I can keep both of these views in the same mental compartment, so to speak; they need not be compartmentalized into different mental slots or thought to apply to different realms or perspectives.)

It is incredibly natural – almost inevitable phenomenologically – to think that I could either go to the movies or to the lecture tonight, that I could either continue working on this essay or take a coffee break, and so forth. It would be jarring to discover that, despite the appearance of the availability of these options, only one path into the future is genuinely available to me. A compatibilist need not come to the one-path conclusion, in the event that theoretical physicists conclusively establish that the conditionals discussed above have associated with them 100 percent probabilities, rather than (say) 99.9 percent probabilities. A compatibilist can embrace the resiliency of this fundamental view of ourselves as agents who (help to) *select* the path the world takes into the future, among various paths it genuinely *could* take. A compatibilist can capture the intuitive idea that a tiny difference (between 100 percent and 99.9 percent) should not make such a huge difference (between our having more than one genuinely available pathway into the future and this pervasive phenomenological fact being just a big delusion). How could such a small change – of the sort envisaged in the probabilities associated with the arcane conditionals of theoretical physics – make such a big difference?

Similarly, it is natural and extraordinarily "basic" for human beings to think of ourselves as (sometimes at least) morally accountable for our choices and behavior. Typically, we think of ourselves as morally responsible precisely in virtue of exercising a distinctive kind of freedom or control; this freedom is traditionally thought to involve exactly the sort of "selection" from among genuinely available alternative possibilities alluded to above. When an agent is morally responsible for his behavior, we typically suppose that he could have (at least at some relevant time) done otherwise.

The assumption that we human beings – most of us, at least – are morally responsible agents (at least sometimes) is extremely important and pervasive. In fact, it is hard to imagine human life without it. (At the very least, such life would be very different from the way we currently understand our

lives – less richly textured and, arguably, not better or more attractive.) A compatibilist need not give up this assumption, even if he were to wake up to the headline, "Causal Determinism is True!" (and he were convinced of its truth). Nor need the compatibilist give up any of his basic metaphysical views – apparently *apriori* metaphysical truths that support his views about free will – simply because the theoretical physicists have established that the relevant probabilities are 100 percent rather than 99 percent. Wouldn't it be bizarre to give up a principle such as that the past is fixed and out of our control or that logical truths are fixed and out of our control, simply because one has been convinced that the probabilities in question are 100 percent rather than 99 percent. A compatibilist need not "flipflop" in this weird and unappealing way.

In ordinary life, and in our moral principles and legal system, we distinguish individuals who behave freely from those who do not. Sam is a "normal" adult human being, who grew up in favorable circumstances (roughly those described in the American TV series, *Leave It To Beaver*). He has no unusual neurophysiological or psychological anomalies or disorders, and he is not in a context in which he is manipulated, brainwashed, coerced, or otherwise "compelled" to do what he does. More specifically, no factors that uncontroversially function to undermine, distort, or thwart the normal human faculty of practical reasoning or execution of the outputs of such reasoning are present. He deliberates in the "normal way" about whether to deliberately withhold pertinent information on his income tax forms, and, although he knows it is morally wrong, he decides to withhold the information and cheat on his taxes anyway.

According to our commonsense way of looking at the world and even our more theoretical moral and legal perspectives, Sam freely chooses to cheat on his income taxes and freely cheats. It is plausible that, given the assumptions I have sketched, he was free just prior to his decision and action *not* to so decide and behave. Insofar as Sam selected his own path, he acted freely and can be held both morally and legally accountable for cheating on his taxes.

On the other hand, we tend to exempt certain agents from *any* moral responsibility in virtue of their lacking even the *capacity* to control their choices and actions; we take it that such individuals are so impaired in their cognitive and/or executive capacities that they cannot *freely* select their path into the future (even if various paths present themselves as genuinely available). Such agents may have significant brain damage or neurological or psychological disorders in virtue of which they are not even capable of exercising the distinctive human capacity of control in any morally significant context. Other agents may have the basic features that underwrite this capacity, but they nevertheless are locally rather than globally exempt from moral responsibility.

On the commonsense view, even as structured and refined by moral and legal analysis, agents who are brainwashed (without their consent), involuntarily subjected to hypnosis or subliminal advertising or other forms of behavioral conditioning, or even direct stimulation of the brain, are not morally responsible for the relevant behavior. But they may be morally responsible for choices and actions that are not the result of these "stock" examples of freedom-undermining and thus responsibility-undermining factors.

Of course, there are difficult cases of significant coercion or pressure that fall short of genuine compulsion, or subliminal suggestion that is influential but not determinative, about which reasonable persons may disagree. Further, there is considerable controversy over the role and significance of early childhood experiences, deprivations, poverty, physical and psychological abuse, and so forth. But even though there are "hard cases," common sense – and moral and legal theory – has it that there are clear cases of freedom and responsibility, on the one hand, and clear cases of the lack of it, on the other.

A compatibilist can maintain this distinction, even if it turns out that the physicists convince us that the probabilities associated with the relevant conditionals – the conditionals linking the past and laws with the present in physics – are 100 percent, rather than 99.9 percent. And this is a significant and attractive feature of compatibilism. Incompatibilism would seem to lead to a collapse of the important distinction between agents such as Sam and thoroughly manipulated or brainwashed or coerced agents. A compatibilist need not deny what seems so obvious, even if the conditionals have attached to them probabilities of 100 percent: there is an important difference between agents such as Sam, who act freely and can be held morally responsible, and individuals who are completely or partially exempt from moral responsibility in virtue of *special* hindrances and disabilities that impair their functioning. Again, a compatibilist's view of human beings as (sometimes) both free and morally responsible agents is *resilient* to the particular empirical discovery that causal determinism is true. Wouldn't it be bizarre if our basic view of ourselves as free and morally responsible, and our distinction between responsible agents and those who are insane or literally unable to control their behavior, would hang on whether the probabilities of the conditionals are 99.9 percent or 100 percent? Again, how can such a tiny change make such a monumental difference?

## 3 A Compatibilist Account of Freedom

One might distinguish between the forward-looking aspects of agency, including practical reasoning, planning, and deliberation, and the backward-looking aspects of agency, including accountability and moral (and legal)

responsibility. I have noted that it is extremely natural and plausible – almost inevitable – to think of ourselves as (sometimes at least) having more than one path branching into the future. This same assumption appears to frame both our deliberation and (retrospectively) our attributions of responsibility. I shall here take it that the possibilities in question are the *same* in both forward-looking and backward-looking aspects of agency: when we deliberate, we naturally presuppose that we have different paths into the future, and when we assign responsibility, we suppose (typically) that the relevant agent had available to him a different path.

In both forward-looking and backward-looking contexts, it is appealing to suppose that the relevant sort of possibility or freedom is analyzed as a certain sort of choice-dependence. That is, when I'm deliberating, it is plausible to suppose that I genuinely can do whatever it is that I would do, if I were so to act: I can go to the movies later insofar as I would go to the movies, if I were to choose to go to the movies, and I can go to the lecture insofar as I would go to the lecture, if I were to choose to go to the lecture, and so forth. On this view, I can do, in the relevant sense of "can," whatever is a (suitable) function of my "will" or choices: the scope of my deliberation about the future is the set of paths along which my behavior is a function of my choices. I do not deliberate about whether to jump to the moon, because (in part at least) I would not successfully jump to the moon, even if I were to choose to jump to the moon.

Similarly, given the assumption of the unity of forward-looking and backward-looking features of agency, the alternative possibilities pertinent to the attribution of responsibility are understood in terms of choice-dependence. That is, on this approach an agent is morally responsible for a certain action only if he could have done otherwise, and he could have done otherwise just in case he would have done otherwise, if he had chosen to do otherwise.

This compatibilist analysis of freedom (or the distinctive sort of possibility relevant to deliberation and responsibility) is called the "conditional analysis" because it suggests that our freedom can be understood in terms of certain conditional statements ("if – then" statements). More specifically, the conditional analysis commends to us the view that an agent *S*'s freedom to do *X* can be understood in terms of the truth of a statement such as, "If *S* were to choose (will, decide, and so forth) to do *X*, *S* would do *X*." The subjunctive conditional specifies the relevant notion of "dependence." The analysis seems to capture important elements of our intuitive picture of what is within the legitimate scope of our deliberation and planning for the future. It also helps to sort out at least some cases in which agents are not morally responsible for their behavior – and to distinguish these from cases in which agents are responsible. If someone is kidnapped and chained, he is presumably not morally responsible for not helping someone in distress insofar as he would

still be in chains (and thus would not succeed in helping), even if he were to choose (decide, will) to help.

But despite its considerable attractions, the conditional analysis, as presented thus far, has fatal problems – problems that should be seen to be fatal even by the compatibilist. First, note that it may be that some outcome is choice-dependent in the way specified by the conditional analysis, and yet there may be some factor that (uncontroversially) impairs or hinders the relevant agent's capacity for choice (in the circumstances in question). This factor could render the agent powerless (in the sense presumably relevant to moral responsibility), even though the outcome is choice-dependent.

So consider the following example due to Keith Lehrer. As a boy, Thomas had a terrible and traumatic experience with a snake. He thus has a pathological aversion to snakes that renders him psychologically incapable of bringing himself to choose to touch a snake (much less pick one up), even as an adult. A snake is in a basket right in front of Thomas. Whereas it is true that *if* Thomas were to choose to pick up the snake, he would do so, it is nevertheless true that Thomas cannot choose to pick up the snake. Intuitively, Thomas cannot pick up the snake – and yet the conditional analysis would have it that he can (in the relevant sense). This is a problem that even a compatibilist should see as significant; the problem clearly does not come from causal determination *per se*. The general form of the problem is that the relevant subjunctive conditional can be true consistently with the actual operation of some factor that intuitively (and apart from any contentious views about the compatibility of causal determinism and freedom) makes it the case that the agent is psychologically incapable of choosing (the act in question) and thus unable to perform the act. Factors that would seem to render an agent psychologically incapable of choice (and which could be seen to do so even by a compatibilist) might include past trauma, subliminal advertising, aversive conditioning, hypnosis, and even direct electric stimulation of the brain.

To make the point starkly, an individual could have his brain directly manipulated (without his consent) so as to choose *Y*. This would presumably render it true that he cannot do *X*, even though it might well be the case that *if* he had chosen to do *X*, he would have done *X*. (Of course, if the individual were to choose *X*, then he would not have been subject to the actual manipulation to which he has been subjected – manipulation that issues in his choosing *Y*.) Think of a demonic (or even well-intentioned) neuroscientist (or even a team including very nice neurophilosophers and neuroscientists!) who can manipulate parts of the brain by using (say) a laser; metaphorically, the laser beam can be thought of as a line that comes from some external source – some source physically external to the individual who is being manipulated and out of his control – at the "end" of which is the individual's choice to do *Y*. The neuroscientist knows the systematic workings of the brain so that she knows

what sort of laser-induced manipulation is bound to produce a choice to do *Y*. Under such circumstances, it would seem ludicrous to suppose that the individual is free to do *X*; and yet it may well be true that *if* he were to choose *X*, the neuroscientist would not have intervened and the individual would successfully do *X*. The outcome is choice-dependent, but the individual is clearly powerless (in the relevant sense).

Some compatibilists about freedom and causal determinism have given up on the conditional analysis in light of such difficulties. Others have sought to give a more refined conditional analysis. So we might distinguish between the generally discredited "simple" conditional analysis, and what might be called the "refined" conditional analysis. Different philosophers have suggested different ways of refining the simple analysis, but the basic idea is somehow to rule out the factors that uncontroversially (that is, without making any assumptions that are contentious within the context of an evaluation of the compatibility of causal determinism and freedom) render an agent unable to choose (and thus unable to act). Along these lines, one might try something like this: An agent *S* can do *X* just in case (i) if *S* were to choose to do *X*, *S* would do *X*, and (ii) the agent is not subject to clandestine hypnosis, subliminal advertising, psychological compulsion resulting from past traumatic experiences, direct stimulation of the brain, neurological damage due to a fall or accident, and so forth . . .

An obvious problem with the refined analysis is the "and so forth . . ." It would seem that an indefinitely large number of other conditions (apparently heterogeneous in nature) could in principle be thought to issue in the relevant sort of incapacity. Additionally, there should be a certain discomfort in countenancing as part of the analysis a list of disparate items with no explanation of what ties them together as a class; from a philosophical point of view, condition (ii) posits an unseemly miscellany. How could one evaluate a proposed addition to the list in a principled way?

Perhaps the compatibilist could simply admit these problems and revise condition (ii) in the following way: (ii') the agent is not subject to *any* factor that would uncontroversially (that is, without making any assumptions that are contentious within the context of an evaluation of the compatibility of causal determinism and freedom) render an agent unable to choose the act in question (and thus unable to act). Despite the obvious problems of incompleteness in this analysis, it might capture something useful; it might capture much of what a compatibilist means by the relevant notion of freedom.

Unfortunately, the proposed revision renders the analysis completely useless in seeking to resolve the controversial issue of whether causal determinism is compatible with freedom (in the relevant sense). This is because all actual choices will be the result of a causally deterministic sequence, if causal determinism is true. Since causal determination obviously is *not* a factor

that uncontroversially renders an agent unable to choose (and thus unable to act), the revised analysis will have results congenial to compatibilism. But it would be dialectically unfair to the incompatibilist to suppose that the revised analysis can be seen uncontroversially to be acceptable; that is, it would clearly beg the question against the incompatibilist to contend that the refined conditional analysis can be seen to be a plausible analysis of the sort of freedom that is under consideration.

To see this more clearly, recall again the metaphor of the line from the neuroscientist to the individual's choice. When the neuroscientist uses clandestine and unconsented-to manipulation by a laser beam, it is plausible that the individual could not have chosen otherwise. Note that, under the circumstances, the only way the individual would have chosen otherwise (would have chosen *X* instead of *Y*) would have been if the neuroscientist had not employed the laser beam as she actually did; that is, it was a necessary condition of the individual's choosing otherwise that the line not have been present, as it were. But the incompatibilist will ask how exactly causal determination of the choice to do *Y* is any different from the neuroscientist's employment of her laser beam to manipulatively induce a choice to do *Y*. The laser beam is not experienced as coercive – it is by hypothesis not experienced at all. The laser is a "subtle" and phenomenologically inaccessible influence that starts entirely "outside" the agent (physically external to the agent and not within his control) and issues (via a process over which the agent has no control) in a choice to do *Y*. But if causal determinism is true, then there is *some* causally deterministic sequence that starts entirely "outside" the agent (physically external to the agent and not within his control) and issues (via a deterministic process) in a choice to do *Y*. The incompatibilist will legitimately ask: what exactly is the difference between the laser beam and the causally deterministic sequence? Metaphorically, they are both lines that start outside the agent and end with the same choice, and both lines must be "erased" to get a different choice.

Now a compatibilist might grant this, but insist on a crucial point – that not all causal sequences are "created equal." More specifically, the compatibilist wishes to insist that not all causally deterministic sequences undermine freedom; a straightforward and "upfront" commitment of the compatibilist is to the idea that we can distinguish among causally deterministic sequences, and, more specifically, that we can distinguish those that involve "compulsion" (or some freedom- and responsibility-undermining factor) from those that do not. Returning to the pictorial metaphor: it may well be that an individual cannot choose or do otherwise when being manipulated by the neuroscientist's laser beam – in this case the individual cannot erase the line from the neuroscientist to his choice. On the other hand, according to the compatibilist, there is no reason to suppose that I am not free either to go to the lecture or go to

the movies later. Even if I do in fact go to the movies, there is no reason – no reason stemming merely from the truth of causal determinism – to suppose that my behavior is causally determined in a *special way* – a way uncontroversially recognized to rule out freedom and responsibility. Thus, even if I do in fact go to the movies later, the compatibilist might say that there is no reason (stemming merely from the truth of causal determinism) to think that I will not be free (at the relevant time) to go to the lecture instead – and thus no reason to suppose that I will not be free (at the relevant time) to erase the line that does in fact connect the past to my choice to go to the movies.

## 4 The Consequence Argument

Yes, it is a basic commitment of the compatibilist – to which I promise we will return below – that not all causally deterministic sequences undermine freedom equally. But there is nevertheless an argument that presents a significant challenge to this commitment and also to the commonsense idea that we can be confident in distinguishing cases of freedom and responsibility from cases where some freedom- and responsibility-undermining factor operates. This argument is a "skeptical argument," rather like the skeptical argument from the possibility of illusion to the conclusion that we don't know what we ordinarily take ourselves to know about the external world. The skeptical argument in epistemology employs basic ingredients of commonsense to challenge other parts of commonsense; that is, it employs ordinary ideas about the possibility of illusion and the concept of knowledge (putatively) to generate the intuitively jarring result that we don't know what we take ourselves to know about the external world. Similarly, the skeptical argument about our freedom employs ordinary ideas about the fixity of the past and the fixity of the natural laws (putatively) to generate the intuitively jarring result that we are not ever free, if causal determinism turns out to be true (something we can't rule out *apriori*). If this skeptical argument is sound, it calls into question *any* compatibilist analysis of freedom (that is, freedom of the sort under consideration – involving the capacity for selection among open alternatives). If the argument is sound, then not only both the simple and refined conditional analysis, but *any* compatibilist analysis (of the relevant sort of freedom) must be rejected. It is thus an extremely powerful and disturbing argument. I think that any honest and serious discussion of compatibilism must address this argument, to which I turn now.

The skeptical argument has been around in one form or another for a very long time. Actually, a structurally similar argument was originally presented thousands of years ago; then the worry was fatalism (more specifically, the idea that the truth values of statements about the future must be fixed and

thus we lack freedom). Then in the Middle Ages the worry stemmed from the doctrine of God's essential omniscience. In the Modern era, our attention has focused primarily (although by no means exclusively) on the threat posed by science – more specifically, the possibility that causal determinism is true. At this point, we simply do not know whether causal determinism is true or not. If it turns out to be true, then all our behavior could in principle be deduced from a complete description of the past and laws of nature. Alternatively, if causal determinism is true, then true propositions about the past and propositions that express the laws of nature entail all our current and future choices and actions.

Here's the argument (very informally). Suppose that causal determinism is indeed true. Given the definition of causal determinism, it follows that my current choice to continue typing (and not take an admittedly much-needed coffee break) is entailed by true propositions about the past and laws of nature. Thus, if I were free (just prior to my actual choice) to choose (and subsequently do) otherwise, then I must have been free so to behave that the past would have been different or the natural laws would have been different. But intuitively the past is "fixed" and out of my control and so are the natural laws. I cannot now do anything that is such that, if I were to do it, the past would have been different (say, John F. Kennedy never would have been assassinated) or the natural laws would be different (say, some things would travel faster than the speed of light [if I've got the natural law in question correct!]). It appears to follow that, despite the natural and almost ineluctable sense I have that I am (sometimes, at least) free to choose and do otherwise, I am never free to choose and do otherwise, if causal determinism obtains.

Although the compatibilist wishes to say that not all causally deterministic sequences equally threaten freedom, the Consequence Argument – so-called by Peter van Inwagen because under causal determinism all our behavior is the consequence of the past plus the laws of nature – appears to imply that causal determinism *per se* rules out the relevant sort of freedom. If the Consequence Argument is sound – and it relies on intuitively plausible ingredients, such as the fixity of the past and natural laws – then the commonsense distinction between cases of "compulsion" and ordinary cases in which freedom is present would vanish, if causal determinism were true; and since we do not know that causal determinism is false, our basic views about ourselves (as free and morally responsible agents) would be called into question.

Recall that it would be uncontroversial that I would not be morally responsible if I were subjected to clandestine (and unconsented-to) manipulation by a neuroscientist's laser beam. As I said above, in such a context, in order for

a different choice to have occurred, the laser beam must not have connected the neurosurgeon with my actual choice. Similarly, if causal determinism were true, then the Consequence Argument brings out the fact that – even in the most "ordinary" circumstances (that is, in the absence of anything that would uncontroversially constitute compulsion) – in order for a different choice to have occurred, the past or the natural laws would have had to have been different: the "line" connecting the past to my choice (via the laws) would have had to have been broken (or erased). The line posited by causal determination appears to be equivalent to the laser beam.

Another way to look at the ingredients that go into the Consequence Argument is to consider the intuitive idea that (as Carl Ginet puts the point) my freedom now is the freedom to add to the given past, holding fixed the laws of nature. In terms of our metaphor, my freedom (on this view) is the freedom to draw a line that *extends* the line that connects the actual past with the present (holding fixed the natural laws). The future may well be a garden of forking paths (in Borges' lovely phrase), but the forking paths all branch off a single line (presumably). The Consequence Argument throws into relief an intuitively jarring implication of compatibilism: the compatibilist cannot embrace the almost undeniable picture of our freedom as the freedom to add to the past, given the laws. If the past is a set of dots, then for our freedom truly to be understandable, we must be able to connect the dots – in more ways than one. Some have said that responsibility involves "making a connection" with values of a certain sort, or "tracking values" in a certain way; but there is even a more fundamental way in which our freedom involves making a connection: we must be able to connect our current actions with our past (holding the natural laws fixed).

In my opinion, the Consequence Argument is a powerful and highly plausible argument. On the other hand, it certainly falls short of being indisputably sound. Some compatibilists – Multiple-Pasts Compatibilists – are willing to say that we can sometimes so act that the past would have been different from what it actually was; alternatively, these compatibilists say that our freedom need not be construed as the freedom to extend the given past, holding the natural laws fixed. On such a view, I might have access to a possibility with a different past associated with it (a possible world with a different past from the actual past) insofar as there are no special "obstacles" in the actual course of events (or the actual world) that "block" such access. Other compatibilists – Local-Miracle Compatibilists – are willing to say that we can sometimes so act that a natural law that actually obtains would not have obtained; some such compatibilists are also willing to countenance small changes in the past as well as the laws. On this sort of view, I might have access to a possibility (or possible world) with slightly different natural laws

from those that obtain actually, as long as these alternative scenarios do not involve widespread and big changes in the laws. This view is defended in a classic paper by David Lewis.

So there is room for compatibilism about causal determinism and the sort of freedom that involves genuine access to alternative possibilities, even in light of the Consequence Argument; excellent philosophers have opted for some response to the Consequence Argument (such as Local-Miracle or Multiple-Pasts Compatibilism). As I said above, the Consequence Argument is not indisputably sound, and thus there is no knockdown argument available that the responses are inadequate.

I do in fact find the Consequence Argument highly plausible, and I am inclined to accept its soundness. I thus think it is important to argue that there is an attractive kind of compatibilism that is indeed consistent with accepting the Consequence Argument as sound. The doctrine of semicompatibilism is the claim that causal determinism is compatible with moral responsibility, quite apart from whether causal determinism rules out the sort of freedom that involves access to alternative possibilities. Note that semicompatibilism in itself does not take a stand on whether the Consequence Argument is sound; it is consistent with acceptance or rejection of the Consequence Argument. My main goal is to defend semicompatibilism, although I am also inclined to accept the soundness of the Consequence Argument. The total package of views I am inclined to accept includes more than semicompatibilism, but semicompatibilism is the principle doctrine I seek to defend here.

## 5 Semicompatibilism and the Frankfurt-examples

Let's say you are driving your car and it is functioning normally. You want to go to the coffee house, so you guide the car to the right (into the parking lot for the coffee house). Your choice to go to the coffee house is based on your own reasons in the normal way, and the car's steering apparatus functions normally. Here you have a certain distinctive kind of control of the car's movements – you have "guidance control" of the car's going to the right. This is more than mere causation or even causal determination; you might have causally determined the car's going to the right by sneezing (and thus jerking the steering wheel to the right) or having an epileptic seizure (and thus slumping over the wheel and causing it to turn to the right) without having exercised this specific and distinctive sort of *control*. Supposing that there are no "special" factors at work – that is, no special psychological impairments, brain lesions, neurological disorders, causal determination, and so forth – and imagining (as above) that the car's steering apparatus is not broken, you had

it in your power (just prior to your actual decision to turn to the right) to continue going straight ahead, or to turn the car to the left, and so forth. That is, although you exercise guidance control in turning the car to the right, you presumably (and apart from special assumptions) possessed freedom to choose and do otherwise: you had "regulative control" over the car's movements. In the normal case, we assume that agents have both guidance and regulative control – a signature sort of control of the car's movements, as well as a characteristic kind of control *over* the car's movements.

Whereas these two sorts of control are typically presumed to go together, they can be prized apart. Suppose that everything is as above, but that the steering apparatus of your car is broken in such a way that, if you had tried to guide the car in any direction other than the one in which you actually guide it, it would have gone to the right anyway – in just the trajectory it actually traveled. The defect in the steering apparatus plays no role in the actual sequence of events, but it would have played a role in the alternative scenario (or range of such scenarios). Given this sort of preemptive overdetermination, although you exhibit guidance control of the car's going to the right, you do *not* have regulative control over the car's movements: it would have gone in precisely the same way, no matter what you were to choose or try.

Of course, in this context you *do* possess *some* regulative control: you could have chosen otherwise, and you could have tried to guide the car in some other direction. This is reminiscent of John Locke's famous example in his *Essay Concerning Human Understanding*. Here a man is transported into a room while he is asleep. When the man awakens, he considers leaving, but he decides to stay in the room for his own reasons. Locke says he voluntarily chooses to stay in the room and voluntarily stays in the room. Unbeknownst to the man, the door to the room is locked, and thus he could not have left the room. According to Locke, the man voluntarily stays in the room, although he does not have the power to leave the room. He exhibits a certain sort of control of his staying in the room (what I would call guidance control), even though he cannot do otherwise than stay in the room (and thus he lacks regulative control over staying in the room). But note that, as in the second car example above, the man could have chosen to leave the room, and he could have tried to do so.

Can structurally similar examples be given in which there is guidance control but *no* regulative control? This is where the "Frankfurt-examples" come in. The contemporary philosopher, Harry Frankfurt, has sought to provide just such an example. One could say that he seeks to put the locked door inside the mind (in terms of Locke's example). In Locke's example, some factor (the locked door) plays no role in the individual's deliberations or choice, and yet its presence renders it true that the individual could not have

done otherwise (could not have left the room). Frankfurt posits some factor that has a similar function in the context of the agent's mind: it plays no role in the agent's actual deliberations or choice, and yet its presence (allegedly) renders it true that the individual could not have chosen otherwise (or done otherwise). If Frankfurt's examples work, then one could in principle entirely prize apart guidance control from regulative control.

Here is my favorite version of a Frankfurt-type case. Jones has left his political decision until the last moment, just as some diners leave their decision about what to order at a restaurant to the moment when the waiter turns to them. In any case, Jones goes into the voting booth, deliberates in the "normal" way, and chooses to vote for the Democrat. On the basis of this choice, Jones votes for the Democrat. Unbeknownst to Jones, he has a chip in his brain that allows a very nice and highly progressive neurosurgeon (Black) to monitor his brain. The neurosurgeon wants Jones to vote for the Democrat, and if she sees that Jones is about to do so, she does not intervene in any way – she merely monitors the brain. If, on the other hand, the neurosurgeon sees that Jones is about to choose to vote for the Republican, she swings into action with her nifty electronic probe and stimulates Jones' brain in such a way as to ensure that he chooses to vote for the Democrat (and goes ahead and votes for the Democrat). Given the set-up, it seems that Jones freely chooses to vote for the Democrat and freely votes for the Democrat, although he could not have chosen or done otherwise: it seems that Jones exhibits guidance control of his vote, but he lacks regulative control over his choice and also his vote. The neurosurgeon's chip and electronic device has brought Locke's locked door into the mind. Just as the locked door plays no role in Locke's man's choice or behavior but nevertheless renders it true that he could not have done otherwise, Black's set-up plays no role in Jones" actual choice or behavior, but it apparently renders it true that he could not have chosen or done otherwise.

How exactly does the neurosurgeon's device reliably know how Jones is about to vote? Frankfurt himself is vague about this point, but let us imagine that Jones's brain registers a certain neurological pattern if Jones is about to choose to vote for the Democrat, and a different pattern if Jones is about to choose to vote Republican. The chip can subtly convey this information to the neurosurgeon, which she can then use to good effect. Of course, the mere possibility of exhibiting a certain neurological pattern is not sufficiently robust to ground ascriptions of moral responsibility, on the picture that requires access to alternative possibilities. That is, if moral responsibility requires the sort of control that involves selection from among various paths that are genuinely open to an agent, the mere possibility of involuntarily exhibiting a certain neurological pattern would not seem to count as the relevant sort of "selection." Put slightly differently, just as it is not enough to

secure moral responsibility that a different choice could have *randomly* occurred, it does not seem to be enough to secure moral responsibility that a different neurological pattern could have been exhibited *involuntarily*. Such an exiguous possibility is a mere "flicker of freedom" and not sufficiently robust to ground moral responsibility, given the picture that requires regulative control for moral responsibility. How could something as important as moral responsibility come from something so thin – and something entirely involuntary?

It is tempting then to suppose that one could have a genuine kind of control – guidance control – that can be entirely prized apart from regulative control and that such control is all the freedom required for moral responsibility. So even if the Consequence Argument were valid and thus all casually deterministic sequences were equally potent in ruling out the sort of control that requires access to alternative possibilities (regulative control), it would *not* follow that all causally deterministic sequences equally threaten guidance control and moral responsibility. That is, if moral responsibility does not require the sort of control that involves access to alternative possibilities, then this opens the possibility of defending a kind of compatibilism, even granting the soundness of the Consequence Argument.

But various philosophers have resisted the temptation to suppose that one can expunge all vestiges of regulative control while at the same time preserving guidance control. They have pointed out that the appearance that Frankfurt-type cases can help to separate guidance from (all) regulative control may be misleading. Perhaps the most illuminating way to put their argument is in terms of a dilemma. The first horn assumes that indeterminism is true in the Frankfurt-examples; in particular, it assumes that the relationship between the "prior sign" (read by the progressive neurosurgeon, Black) is causally indeterministic. It follows that right up until the time Jones begins to choose, he can begin to choose otherwise; after all, the prior sign (together with other factors) fall short of causally determining the actual choice. Thus, there emerges a robust alternative possibility – the possibility of beginning to choose otherwise. This is no mere flicker of freedom; although it may be blocked or thwarted before it is completed or comes to fruition, it is nevertheless a voluntary episode – the initiation of choice.

The second horn of the dilemma assumes that causal determinism is true in the examples. Given the assumption of causal determination, it would appear to be straightforwardly question-begging to say that Jones is obviously morally responsible for his choice and behavior (despite lacking genuine access to alternative possibilities). After all, this (the compatibility of causal determination and moral responsibility) is precisely what is at issue!

The dilemma is powerful, but I am not convinced that it presents an insuperable objection to the employment of the Frankfurt-examples as part

of a general strategy of defending compatibilism; as a matter of fact, I reject it. First, there have been various attempts at providing explicitly indeterministic versions of the Frankfurt-cases, and I think some of them are promising. I don't think it is obvious that one could not construct a Frankfurt-example (under the explicit assumption of causal indeterminism) in which there are *no* robust alternative possibilities. Recall that it is not enough for the proponent of the regulative control requirement to identify just any sort of alternative possibility; rather, he needs to find an alternative possibility that is sufficiently *robust* to ground attributions of moral responsibility, given the regulative control picture. If the ground of moral responsibility is a certain sort of *selection* from genuinely available paths into the future, then paths with mere accidental or arbitrary events would seem to be irrelevant. This is the idea of the irrelevance of exiguous alternatives. In my view, to seek to get responsibility out of mere flickers of freedom is akin to alchemy. Here I am in agreement with the libertarian, Robert Kane, who insists on the "dual-voluntariness" constraint on moral responsibility.

On the second horn of the dilemma, causal determinism is assumed. I agree that it would be dialectically rash to conclude precipitously from mere inspection of the example presented above that Jones is morally responsible for his choice and voting behavior. Rather, my approach would be more circumspect. First, I would note that the distinctive contribution of the Frankfurt-examples is to suggest that if Jones is not morally responsible for his choice and behavior, this is *not* because he lacks genuine access to (robust) alternative possibilities. After all, in the example Black's set-up is sufficient for Jones's choosing and acting as he actually does, but intuitively it is *irrelevant* to Jones's moral responsibility. That is, we can identify a factor – Black's elaborate set-up – that is (perhaps in conjunction with other features of the example) sufficient for Jones's actual kind of choice and behavior, but it plays no actual role in Jones's deliberations or actions; Black's set-up could have been subtracted from the situation and the actual sequence would have flowed in exactly the way it actually did. When something is in this way irrelevant to what happens in the actual sequence issuing in an agent's choice and behavior, it would seem to be irrelevant to his moral responsibility.

So the distinctive element added by the Frankfurt-type examples, under the assumption of causal determinisim, is this: if the relevant agent is not morally responsible, it is not because of his lack of regulative control. Alternatively, we could say that they show that it is not the lack of genuine access to alternative possibilities (regulative control) in itself (and apart from pointing to other factors) that rules out moral responsibility. Now we can ask whether there is some other factor – some factor that plays a role in the actual sequence – that rules out moral responsibility, if causal determinism obtains. We will turn to a more thorough and careful consideration of such "actual-

sequence factors" below. For now, I simply wish to note that there is nothing question-begging or dialectically inappropriate about how I have invoked the Frankfurt-examples thus far (on the second horn), and their distinctive role is to call into question the relevance or importance of regulative control in grounding moral responsibility (in the way presented described above).

Taking stock, in section 4 I presented the Consequence Argument. I did not officially endorse its conclusion, although I am inclined to believe that the argument is valid and, further, that its premises are based on extremely plausible ingredients. I suggested that it would be prudent to seek a defense of compatibilism that does not presuppose that the Consequence Argument is unsound. Here I have presented the rudimentary first steps toward the elaboration of just such a compatibilism. I have invoked the Frankfurt-examples (the prototypes of which are in John Locke) to support the contention that moral responsibility does not require regulative control (or the sort of freedom that involves genuine access to alternative possibilities), but only guidance control. Further, I have suggested that (thus far at least) there is no reason to suppose that causal determinism is inconsistent with guidance control. Better: I have contended that even if causal determinism threatens regulative control, it does not thereby threaten guidance control. In the next section I will explore whether there are other reasons (apart from regulative control) in virtue of which causal determinism is incompatible with moral responsibility.

## 6 Source Incompatibilism

In section 5 I sought to provide a sketch of an argument for the conclusion that regulative control (and genuine access to alternative possibilities) is not required for moral responsibility. I suggested that alternative possibilities do not – in themselves and apart from indicating something else – ground ascriptions of moral responsibility. Note that if causal determinism is true and the Consequence Argument is sound, then there are no alternative possibilities available to agents – even mere flickers of freedom. So one reason someone might insist on the importance of even exiguous possibilities – mere flickers of freedom – would be as *indicators* of something in the actual sequence – the lack of causal determination. Some philosophers have argued then that even though access to alternative possibilities does not in itself explain or ground moral responsibility ascriptions, it is a necessary condition of such attributions. This view about the nature of the relevance (or importance) of access to alternative possibilities stems from a more fundamental idea that causal determination in itself rules out moral responsibility (apart from threatening access to alternative possibilities).

Why might one think that causal determinism rules out moral responsibility "directly" (and not in virtue of closing off alternative pathways)? We might think of it this way – again in terms of the metaphor of a line. As above, suppose that we represent causal determination as a line from some features of the past to one's current choice (and behavior). Above I pointed out that a compatibilist about causal determinism and the sort of control that involves genuine access to alternative possibilities must think that agents can sometimes sever or erase the line that actually connects the dots in the past and present. Here we are not supposing that an agent can have that sort of freedom; rather, we are considering whether the actual-sequence line must be severed or broken, in order for moral responsibility to be plausibly ascribed to the agent. That is, here we are not wondering about the agent's power to break or erase the line; here we are considering whether we as theorists must posit a broken or dotted line in order for it to be plausible that the agent is morally responsible. In terms of the metaphor, the question is about the significance of the gaps. Why must there be gaps or spaces between the dots, in order to make room for moral responsibility?

Of course, one reason it might have been supposed that we need to posit gaps or spaces is to allow for access to alternative possibilities. Here the gaps give rise to "elbow room" that provides the space to pursue alternative paths. To employ another metaphor (with which I am, lamentably, well-acquainted, living in Southern California): if traffic is proceeding on the freeway literally "bumper-to-bumper," there are no gaps or spaces between the cars. Given this, no car after the first car can exit the freeway or the lane in which it is traveling; the lack of space between cars renders it impossible for the drivers to change directions (although of course they can try). But the above argumentation (pertaining to the Frankfurt-cases) showed that the significance of the gaps or spaces in the actual sequence cannot be to make room for access to alternative possibilities. We must look for some other significance of the gaps.

Admittedly, driving (especially rapidly!) under the envisaged circumstances would be very uncomfortable. But – leaving aside elbow room for changes of direction – why might we insist on not being "scrunched up" in the indicated way? In this section I shall explore three salient views about the (intrinsic) importance of the gaps. First, it might be suggested that without gaps, there is no room for "agents." Second, I shall consider the related idea that only with spaces can one have "activity," rather than mere passivity. Finally, I shall consider the notion that if the line is entirely filled in, then the "source" of the relevant choices and behavior must be external to the agent in a way that rules out moral responsibility.

It is sometimes proclaimed that if causal determinism were true, there would be no room for "agency." As with most propositions enunciated with

so much portentous certainty, it is quite obscure even what the claim is – let alone whether it is true! Here is one way of trying to make sense of the pronouncements. If causal determinism is true, then individual agents – persons or selves – are entirely composed of events (construed broadly) that have (deterministic) causal interaction with the external world. If we are just complexes of events related deterministically to our surroundings, there might not seem to be space for "selves" or "persons." The worry here appears to be that causal determinism entails a kind of "reductionism" about the self or agent – a reductionism alleged to be unattractive and implausible. If all there is is a bunch of events in a deterministic causal network, where, it might be asked, is the self or agent? The self is crowded out.

I find this worry hard to get a handle on. There are various versions of "reductionism" – about meaning, explanation, metaphysics, and so forth. This is not the place to address the various kinds of reductionism, or their relationship to the doctrine of causal determinism. The main point I wish to make is that it is not at all clear that causal determinism entails *any* kind of obviously problematic reductionism. I suppose one might seek to argue that the self cannot be composed entirely of events (broadly construed) that are within a deterministic causal network in nature – but the argument would be at best contentious. In the end, I think that it would not be obvious or clear that causal determinism doesn't leave room for a self or individual agent, in any sense in which we are uncontroversially such selves or agents.

One can be misled by, as it were, looking in the wrong place. If you were to micro-miniaturize yourself and explore the human nervous system from inside the body, all you would see (presumably) are cells bumping into each other. It does not follow that the human nervous system cannot support thought or consciousness. Of course, how exactly consciousness supervenes on or emerges from a physical basis is contentious; indeed, some philosophers argue that the phenomena of consciousness cannot come from a mere material or physical basis, no matter how structurally and functionally intricate. But it is at least important to note that there is vigorous disagreement about this issue; it is not obvious and clear that consciousness cannot have a physical basis (in some relevant sense). Similarly, it is not at all uncontentious that the self cannot be composed of a set of events (broadly construed) located in a causally deterministic niche within nature (or perhaps the external world). (It should be noted here that I do not assume that causal determinism entails physicalism or materialism; I am simply drawing an analogy between discussions of reductionism in the context of consciousness or the mind and agency.)

If one takes apart a television set and looks at the inner works of the TV, all one sees is a bunch of physical components. I suppose one could be

completely perplexed if one tries to find the TV at the level of the components. But of course it does not follow that the TV is not composed of a set of components, perhaps structurally and functionally related in the right way. The TV is nothing more than these parts, structured and capable of functioning in certain characteristic ways; to look for the TV at the level of the parts is to look in the wrong place. Similarly, it may well be that a self or agent is a set of events (broadly construed) within a deterministic causal network structured and capable of functioning in certain distinctive ways; to look for the self or agent at the level of the parts is to look in the wrong place. Further, it is not at all clear how simply breaking the line (representing causal determinism) and inserting spaces really helps us to find the self; if anything, it threatens to make matters worse. (It would obviously be unpromising to look for the TV set in the spaces between the components!) The self might disappear – lost in space.

There may well be something at a deep level of analysis to the worry about causal determinism not leaving room for the self or agent, but the worry is not at all obvious. I do not suppose that I have proved that the worry is not to be taken seriously; rather, I have simply suggested that it is not decisive (at least as thus far developed). The second worry is perhaps closely related to the first. An agent is not simply a coherent and separate individual or self – an agent is "active" in a signature sense, rather than merely or wholly passive. The second worry then is that causal determinism rules out *activity* in the relevant sense – it would render us all completely passive. Harry Frankfurt, who many philosophers assume is a compatibilist, has recently expressed the worry that causal determinism might rule out activity; he thus concludes that we do not yet have a conclusive defense of compatibilism.

I agree that the relevant notion of activity is difficult to put one's finger on, and its relationship to causal determinism is obscure. Not surprisingly, however, I am not at all convinced that causal determinism is inconsistent with the relevant notion of activity. Unfortunately, as above, I cannot offer a decisive answer to the worry about the relationship between causal determinism and being active; instead, I seek to present some considerations that, I believe, at least render it plausible that causal determination is consistent with activity (or perhaps that render it less plausible that causal determination rules out activity, in the relevant sense). Although I agree with Frankfurt that we don't have a satisfactory grasp of the distinction between activity and passivity, I hope at least to assuage some of the anxieties.

Note first that there is a perfectly ordinary and commonplace way of making the distinction between being active and being passive that seems entirely orthogonal to issues pertaining to causal determinism; more specifically, being active in this sense is in no way threatened by causal determination. So, for example, we characterize someone as "active" in a relationship

insofar as she "takes the lead" in various ways: she typically and frequently makes suggestions about activities, projects, and ways of doing things, she anticipates potential problems and seeks to head them off, she does not defer to her partner's wishes or suggestions easily, and, in general, she "listens to her own voice" or perhaps "takes cues from herself." She is a leader, as opposed to a follower. On the other hand, someone is relatively passive insofar as he is deferential: he typically allows his partner to take the lead in setting policies and making suggestions, he tries to please her and often defers to her wishes, and he basically takes his cues from her, rather than listening to his own voice. Of course, this is all very rough, but it should be evident that nothing in the distinction between being "active" in the ordinary interpersonal sense requires the absence of causal determination: taking one's cues from within can be part of a causally deterministic sequence.

Now someone might object that this "ordinary and commonplace" way of making the distinction is *not* what is at issue in the more refined reflection on moral responsibility in which we are currently engaged. Here (it is alleged) we can just see that the ordinary notion of activity is not enough, and in order to be active in the sense that is required for moral responsibility, one's choices, behavior, and the formation of one's character over time cannot be causally determined. I do not know how exactly to respond to this move, except to point out that it is to abandon the idea that there is a widely shared and appealing notion of "being active" – a notion that is embedded in our commonsense ways of understanding ourselves and interpreting our behavior – by reference to which we can see that causal determination would rule out moral responsibility. The relatively widely shared and appealing notion of "being active" is completely consistent with causal determinism. In contrast, the incompatibilist is invoking a rather *special* and *particular* notion of "being active," and it is not at all clear or uncontroversial that *this* notion of activity is required for moral responsibility. Why, one might ask, demand this rather rarified sort of "activity" for moral responsibility?

The dialectic here is similar to what might be called the "dialectic of initiation." There is a widely shared and relatively uncontroversial view that moral responsibility requires that I be the initiator of my behavior – that I "start" the sequence that issues in certain consequences in the world. But, as with activity, there is a perfectly ordinary and commonplace notion of "initiation" that is entirely orthogonal to issues pertaining to causal determinism; initiating something, in this sense, does not require the falsity of causal determinism. A boy may start a fire by lighting a match and throwing it in dry weeds, even in a causally deterministic world. Similarly, the Beatles started the so-called "British Invasion" of the United States involving rock bands in the 1960s and – to invoke a less well-reported phenomenon – Kant started the "transcendental turn" in philosophy, quite apart from whether or

not causal determinism is true. We make perfectly reasonable attributions of initiation without bothering to consider whether causal determinism is true; and, upon reflection, it is not evident that we ought to retract the claims, even under the assumption of causal determinism.

Now an incompatibilist might concede that there is a perfectly reasonable and "ordinary" notion of initiation that is not in fact threatened by causal determinism. But, as above, he might insist that this ordinary notion is not enough for moral responsibility; what is required, on this view, is a kind of initiation that is inconsistent with causal determination. After all, it might seem that the boy does not "really" start the fire, if his choice was causally determined by previous events, and so forth. Again, I do not have a decisive reply to this move. But I would emphasize (as above) that the incompatibilist's strategy here abandons the idea that there is a widely shared and attractive notion of initiation – some notion of initiation that should be appealing to both compatibilists and incompatibilists – that is uncontroverisally ruled out by causal determinism. The relatively widely shared and appealing notion is entirely consistent with causal determinism. In contrast, the incompatibilist is identifying a *special* and arguably *rarified* notion of initiation, and it is not at all clear or uncontroversial that *this* notion of initiation is required for moral responsibility.

Returning to the puzzling notion of "being active," it is almost as if the incompatibilist is in the grip of a certain metaphor – that of a row of dominos. Suppose you impart enough force to the first domino to cause (via a deterministic sequence) the rest of the dominos in the row to topple (one by one). Each step along the way is causally determined. When the second domino falls, it deterministically causes the third domino to fall, and so forth. Clearly, the third domino is entirely passive. But it would be a mistake to suppose that all causally deterministic sequences are relevantly similar to this simple model of falling dominos! Note that the capacity for activity (in the relevant sense) seems to require mental states, including "executive" states, such as volitions, decisions, and choices. A domino has no mental states and thus no executive states of the pertinent kind; a domino is not the sort of thing that can be active, quite apart from whether it is embedded in a causally deterministic sequence.

I now turn to the third worry mentioned above – that causal determinism would entail that an individual would not be the "source" of his choices and actions, in the sense of "sourcehood" required for moral responsibility. The idea that an individual must himself be the source of his behavior – or perhaps that the source must be "internal" to the individual – is similar to the idea that an agent must "initiate" his behavior. Above I pointed out that there is a perfectly reasonable notion of initiation that is completely compatible with causal determination; of course, the same considerations apply to the notion

of "sourcehood." Here I wish to explore some of the reasons why philosophers have contended that causal determinism rules out sourcehood in the sense required by moral responsibility, and I wish to offer some plausibility-arguments against this contention. As above, I do not suppose that I will have exhausted the possible motivations for an incompatibilistic sourcehood requirement, or that I will have offered knockdown arguments against such a requirement. My goal is to lay out some of the salient motivations for the worry that causal determination would threaten sourcehood, and to suggest that the worry may issue from a mistaken picture – an inflated conception of the sort of control we must possess in order to be morally responsible.

If causal determinism is true, then our behavior is the result of causally deterministic sequences that began well before we were even born. Since we are not responsible for initiating these sequences, and since our decisions and behavior are the necessary results of them, we are not "ultimately" in control of our behavior. Saul Smilansky has a nice phrase for what he takes to be the responsibility-undermining characterstic of causal determination: the "mere unfolding of the given." Since our behavior under causal determinism would be the mere unfolding of the given, Smilansky concludes that compatibilism is "morally shallow." The argument seems to be that since we have no control over the beginnings of the causal sequences that causally necessitate our choices and behavior, it follows that from the perspective of our control, there might just as easily have been *different* beginnings of those sequences, and thus different choices and actions. The locus of control would not be "internal" to us, in the required sense; from a perspective that considers the possibility of different beginnings of the sequences, it is entirely arbitrary or a matter of pure luck that we behave as we actually do.

Perhaps the incompatibilistic sourcehood requirement comes from, or is suggested by, a certain picture of agency. On this picture, the locus of control must be entirely *within* us, if we are to be morally responsible. But when there is some factor that is external to us, over which we have no control, and upon which our behavior and even "the way we are" is (or might be) counterfactually dependent, the locus of control is not within us in the relevant way. It is as if the proponent of the incompatibilistic sourcehood constraint thinks of agents who are morally responsible as having "total control." An agent has total control when the locus of control is "within him" in a certain specific way. More specifically, an agent has total control over $X$ only if for any factor $f$ which is a causal contributor to $X$ and which is such that if $f$ were not to occur, then $X$ would (or might) not occur, then $X$ has control over $f$.

But total control is a total fantasy – metaphysical megalomania, if anything is. The sun is shining (through the smog), and its continuing to shine is a contributing causal factor to my continuing to exist, continuing to be an agent, and so forth. If the sun were to flicker out, I would not continue

to exist, continue to be an agent, or engage in any behavior. So the sun's continuing to shine is a contributing cause to my behavior, is completely out of my control, and is such that, if it were not to occur, I would not even exist. Molly Bloom said in James Joyce's *Ulysses*, "The sun shines for you . . ."; and it is a good thing. Obviously, the sun's continuing to shine is just one of an indefinitely large number of such factors: a huge meteorite's not hitting the United States, my not being hit by a lightning bolt, and so forth.

The sun's continuing to shine is a (background) sustaining cause of my existence and agency. Consider now the fact that my parents did not seriously injure me when I was young and helpless. (Now I am old and helpless!) That they took good care of me was a contributing cause of my developing into an agent at all. Had they significantly abused and injured me, I would or at least might not have developed into an agent at all. And of course how my parents treated me when I was an infant was entirely out of my control. Clearly, my parents treating me in a gentle way is just one of an indefinitely large number of such factors: my not falling on my head and incurring a significant brain injury when I was young, my not having been born with a terrible neurological disorder, and so forth. (Note that a factor that is described negatively can be transformed easily into a positively described factor: my not being born with a neurological disorder is my being born with a normal central nervous system, and so forth.)

So total control is a chimaera. It is manifestly ludicrous to aspire to it or to regret its absence. The locus of control is not wholly within us. We do not exist in a protective bubble of control. Rather, we are thoroughly and pervasively subject to luck: actual causal factors entirely out of our control are such that, if they were not to occur, things at least might be very different. Quite apart from any special assumption about causal determinism, we can see that from a broader perspective, it is entirely a matter of luck or arbitrary that I behave as I do (or even that I developed into an agent at all – or have maintained that status). Although it is perfectly reasonable to wish to be the source of one's choices and behavior, it is not reasonable to interpret the relevant notion of sourcehood in terms of total control and internality (understood as above).

Now of course one might seek to motivate an incompatibilistic source requirement in various different ways. But my suggestion is that, once one sees that the picture that favors total control is inflated and illusory, one might have considerably less inclination to accept such a requirement for *any* reason. Of course, there may well be an important difference between our lack of control of external necessary or enabling conditions, and our lack of control of external causally sufficient conditions. I grant this. But my suggestion (and it is merely a suggestion) is that, once one recognizes the pervasiveness of a certain sort of luck, one will find an incompatibilistic source condition less

attractive. After all, there are necessary conditions that enable us to be agents or "set the stage" for our agency and which are entirely out of our control. Given this, one might wonder why it is problematic that there might be causally sufficient conditions for our behavior that are external and entirely out of our control.

Return to our recurrent pictorial device of the line. Imagine here that our agency is represented by a horizontal line-segment from point b to point c. This is the Agency Line. Now suppose there is a vertical line coming from below, with an arrow pointing toward the Agency Line. The vertical line represents a causally necessary condition (or an enabling condition), such as the sun's shining; the sun's shining causally sustains and "sets the stage" for the exercise of agency. Now add a line that is (like the Agency Line) horizontal, starting to the left of point b at some point a, connecting a and b, and with an arrow pointing toward b. Suppose that the relevant agent is not in control of this antecedent causal sequence "pointing horizontally toward b," just as he is not in control of the sun's continuing to shine. My question is: what is the difference between the vertical and horizontal lines? More carefully, if one is not troubled by the existence of the vertical line, why be troubled by the horizontal line? The two lines are equally "external" to the Agency Line, and thus mere appeal to internality will not distinguish the lines. (Of course, this point is consistent with there being some *other* factor – apart from mere internality – by reference to which one could distinguish necessary and sufficient causal conditions in the relevant respect.)

Again, I do not suppose that I have offered a knockdown argument that there can be no relevant difference between the role of causally necessary and causally sufficient conditions. Rather, I have simply presented some considerations that suggest that the difference may not be as deep and important (in this dialectical context) as some have supposed. In fact, the pictorial metaphor of the horizontal and vertical lines is a bit misleading. This is because the sufficient causal condition for any choice or action will presumably be (or be analyzable as) a large conjunction, with individual conjuncts including such enabling factors as the sun's continuing to shine, and so forth. That is, the vertical line (or, perhaps more accurately, lines) is already included in the horizontal line-segment connecting a and b.

But this may further suggest that the mere existence of the horizontal line connecting a and b should not trouble us unduly. After all, each of the many conjuncts that are necessary causal conditions of one's agency does not trouble us; it is unclear why conjoining them and adding some additional ingredients so that we now have a causally sufficient condition should be troubling. Let us simply posit that one finds a causally sufficient condition troubling because being subject to such causation somehow "pushes" us; perhaps this view comes from a certain idea that natural laws "push" or "compel" us. Now, quite apart

from the specific considerations at issue here this does not seem to me to be a plausible view about the necessitating component of natural laws. (Natural laws do necessitate in some way; but it is contentious that they do so by, as it were, "pushing.") And note the implausibility of supposing that, whereas none of the individual necessary conditions (such as the sun's continuing to shine) pushes or compels us, once we add these all together and perhaps add some "triggering" cause (that creates a sufficient condition against the background of the conjunction of enabling necessary condition), we suddenly get "pushing" or "compulsion." So if one does not find the vertical line troubling, why be disturbed by the horizontal line?

Imagine a beautiful dive. The diver may exhibit great skill and even courage. Given that the diver freely engages in the competition, he may be commendable for his dive, even though he had no control over the building of the diving board, or the fact that it is not subtly cracked as a result of a lightning bolt during the previous evening, and so forth. He controls his dive, although he obviously does not control all the causally enabling conditions. His agency takes place literally on a platform that is not his own creation. Further, an agent who dives into a cold river to save a drowning child may control his choice and behavior, and be morally responsible for it, quite apart from issues pertaining to the creation or maintenance of the "platform" from which he leaps.

Nietzsche famously said, "the *causa sui* is the best self-contradiction that has been conceived so far; it is a sort of rape and perversion of logic." The quotation is from *Twilight of the Idols, or: How to Philosophize with a Hammer*, section 8, "The Four Great Errors." Now I'm not sure that the *causa sui* would make my Top Ten List of Good (or perhaps Egregious) Self-Contradictions, but to be the cause of oneself (in a stringent way) is surely an unreasonable aspiration. Whereas some philosophers would claim (with Nietzsche) that being a *causa sui* is both ludicrous and part of commonsense, I would urge that we note that being the "initiator" or "source" of our choices and behavior is indeed part of commonsense, but that it is inchoate and undeveloped in commonsense. We should not be quick to attribute a ludicrous and obviously self-contradictory notion to commonsense. Rather, we should seek to capture the kernel of truth embedded in our ordinary conceptual scheme and articulate it in a more plausible, attractive way.

In this section I have considered the possibility that causal determinism rules out moral responsibility "directly" (and not via threatening genuine access to alternative possibilities). If causal indeterminism issues in spaces or gaps, the question is, "What is the significance of the gaps for moral responsibility, given that what matters about the gaps is not that it provides elbow room for alternative possibilities or changes of direction?" The view that the gaps matter – apart from providing such elbow room – is typically called

source incompatibilism. Here I have considered three versions of the generic doctrine of source incompatibilism, and I have sought to defend compatibilism against these worries. Of course, as with the view that causal determinism rules out moral responsibility indirectly (via threatening regulative control), I have not offered knockdown arguments; but I hope to have at least taken much of the sting out of the objections. That is, I hope (at least) to have shown how compatibilism of a certain sort – semicompatibilism – is left standing after the best punches have been thrown by its opponents.

## 7 Why Be a Semicompatibilist?

I began with some of the salient motivations for being a compatibilist. Especially because our most fundamental views of ourselves as free and morally responsible should not, as it were, "hang on a thread" – should not depend on subtle and arcane deliverances of theoretical physicists – there are strong attractions to compatibilism. But we have also looked at serious objections to compatibilism. First we considered the Consequence Argument. This argument employs extremely plausible ingredients, such as the fixity of the past and natural laws, to derive the conclusion that if causal determinism were to obtain, then no one is free in the sense of having genuine access to alternative possibilities. I have suggested (in section 5) that we distinguish two kinds of freedom or control – regulative control (which requires genuine access to alternative possibilities) and guidance control (which involves a distinctive kind of guidance but not access to alternative possibilities). We can now sidestep the difficulties presented by the Consequence Argument by noting that guidance control – not in any way threatened by the Consequence Argument – is the sort of freedom or control bound up with moral responsibility. Semicompatibilism contends that moral responsibility is compatible with causal determinism, quite apart from whether causal determinism threatens regulative control. (Semicompatibilism is thus consistent with, although it does not in itself *require*, the acceptance of the soundness of the Consequence Argument.) Further, I have argued (in section 6) that some of the most salient versions of Source Incompatibilism are, at best, inconclusive. Given the considerable attractions of compatibilism (all of which are enjoyed by semicompatibilism), I believe that a careful evaluation of the dialectical situation as a whole should issue in an acceptance of semicompatibilism.

But why should one be a semicompatibilist rather than a traditional compatibilist? What exactly is the benefit of switching from a model that requires regulative control to one that only requires guidance control? Some philosophers have claimed that semicompatibilism is merely old wine in new bottles. They have stated that it seems to have a "scholastic air" to it (in Jay Wallace's

phrase), and that no one really cares (or for that matter ever has cared) about a sort of freedom that holds all of the past and laws fixed – regulative control. One important compatibilist – Gary Watson – has pointed out that, given the definition of causal determinism, it is *blatantly obvious* that causal determinism would rule out any kind of freedom that required that all of the past and natural laws be held fixed; Watson wonders why we should belabor this point, or why it should be in the least surprising. For Watson (and others, especially traditional compatibilists), semicompatibilism is not so much an innovation as something always presupposed by traditional compatibilists. Another influential compatibilist, Daniel Dennett, has recently said that when we are interested in whether someone could have done otherwise, we are *never* interested in whether he could have extended the (entire) actual past, holding fixed the natural laws, "unless we are doing philosophy and confronting the [Consequence Argument]."

These remarks puzzle me for various reasons. In our phenomenology as agents, it is quite natural to think that in deliberating about the future, we are selecting from among various options that are genuinely metaphysically open to us. Additionally, our commonsense theorizing about our moral and legal responsibility presupposes that sometimes at least we could have done otherwise. As I stated above, it is natural to *identify* the alternative possibilities presupposed in the forward-looking component of agency (deliberation) and the backward-looking component (moral and legal responsibility).

Certainly, for thousands of years philosophers have wondered about whether we have such freedom – the freedom that involves freedom to select from genuinely open options – in light of such worries as the prior truth values of statements about the future ("fatalism") and God's existence and essential omniscience. The worries about prior truth values of statements about the future and also God's omniscience, construed as foreknowledge, stem precisely from a deep concern we have about the relationship between the past and our freedom. More specifically, the classic debates – that have dominated the discussion of free will for thousands of years – assume that our freedom is exactly the freedom to extend the actual past. Some have pointed out that it is not clear that the *truth* of certain statements about the future is a fact about the past in the same way as paradigmatically fixed facts about the past, and others have said the same thing about God's beliefs about the future. But note that these debates take place within the shared framework of an assumption that indeed our freedom is the freedom to "go from here" – to extend the actual past (defined in terms of paradigmatically [temporally] nonrelational facts) in its entirety; the only question is whether (say) God's prior beliefs are relevantly similar to the standard temporally nonrelational (and thus fixed) facts.

So to suggest that no one has ever seriously worried about whether we have regulative control in our present circumstance, given how the past has led to these circumstances, seems to ignore great swaths of the history of philosophy. In the Modern Era the debates have included worries that stem from science and the possibility of causal determinism, as well as logic and religion. Of course, I do not contend that everyone has agreed that we need regulative control for moral responsibility, or that such control needs to be analyzed in terms of stringent fixity of the past (or laws) constraints; reasonable people can disagree about these matters. But that is not to say that the history of debates about free will has not been replete with disputes about precisely these matters!

Given the history of contentious and apparently intractable debates about the relationship between such doctrines as God's omniscience and regulative control, and also causal determinism and regulative control, it would seem that it would clearly be a helpful (and substantial!) step in the right direction to *sidestep* these debates by developing a defense of compatibilism that does not require their resolution. Indeed, my defense of semicompatiblism allows (although does not require) one to accept that we never have genuine metaphysical access to alternative possibilities.

Perhaps my basic point here is that it can help us make considerable dialectical progress in debates with those inclined toward incompatibilism to *allow* them their views about the fixity of the past and the fixity of the natural laws. Even if one does not oneself care at all about possessing the sort of freedom that involves the power to extend the actual past (in its entirely), holding fixed the natural laws, it is clear that *others* do care about precisely this sort of freedom. In my view, any compatibilist who ignores the natural attractiveness of the desire to have this sort of freedom vitiates his dialectical position significantly and, indeed, unnecessarily. Such a compatibilist risks being dismissed, or at least finding himself in an intractable dialectical stalemate. On the other hand, I can admit that it is natural to think of oneself as possessing regulative control, and that it is plausible to analyze this in terms of the power to add to the actual past (the entirety of the temporally nonrelational past), holding fixed the laws of nature. A semicompatibilist need not dismiss out of hand, or profess puzzlement, about what is surely an intuitively natural set of views – he can embrace the strongest points of the incompatibilist and still defend compatibilism!

I believe that the strongest or most compelling feature in the incompatibilist's arsenal is the Consequence Argument. I believe that the ingredients that go into the Consequence Argument – the fixity of the past and the natural laws – are considerably more gripping than the ingredients of source incompatibilism. Thus, semicompatibilism is able to embrace the most

attractive features of incompatibilism without thereby having to accept its least attractive feature – that our freedom and moral responsibility "hangs on a thread," or, in libertarianism, that we can know from our philosophers' arm-chairs that causal determinism is false. We semicompatibilists can have our cake and eat it too.

In an important challenge, Gary Watson has asked what is gained in terms of securing "control" by positing indeterminism – by severing the line that connects the past to our choices and actions. From a certain perspective, it seems that adding indeterministic gaps or spaces can etiolate our control rather than strengthening it; it now becomes unclear that our choices and actions are really *ours*. A semicompatibilist has an answer to Watson's challenge. For the semicompatibilist, it is important to distinguish the two kinds of control discussed above: guidance control and regulative control. The Consequence Argument appears to show that there can be no regulative control, if causal determinism obtains. So the assumption of indeterminism at least opens the door to the possibility of regulative control – it removes what appears to be an insuperable obstacle to such control, even if other challenges remain. Of course, indeterminism in itself does not appear to help to secure guidance control, unless one accepts source incompatibilism. So the incompatibilist can embrace the kernel of truth behind Watson's challenge – that guidance control is arguably not enhanced by positing indeterminism – while also explaining why it is tempting and natural for an incompatibilist to suppose that positing causal indeterminism can help to secure control: guidance control is not enhanced, but (arguably) the door is opened for regulative control.

My fundamental contention is that semicompatibilism can help by allowing us to sidestep traditionally intractable debates. If one can *grant* that there is an important kind of freedom that is (arguably, at least) ruled out by causal determinism – a notion that is typically and naturally associated with deliberation and moral responsibility – and yet still present a persuasive case for compatibilism, one is at a significant dialectic advantage. The traditional compatibilist has a much more difficult project: he must defeat the Consequence Argument as well as source incompatiblism. Why needlessly make your job more difficult than it already is?

I am inclined to add that I really am puzzled by those who say that it is not natural or plausible to suppose that we (sometimes at least) possess the freedom to choose and do otherwise, where this involves the power to extend the (entire) given past, holding fixed the laws of nature. Typically, of course, we are not in a position to know the entirety of the past or the complete statement of the laws of nature, so we do not seek to find such things out in our ordinary lives. When I deliberate, I don't check the total statement of the past and the laws of nature. But it certainly does *not* follow that I don't

implicitly presuppose that my freedom is the power to extend the actual past, whatever that is, holding fixed the natural laws, whatever they are. Our epistemic limitations imply that we do not worry about compatibility with the past and the laws when we plan for the future; but this is perfectly compatible with a background presupposition that whatever we can do must be connected to the past by a line (holding fixed the natural laws).

Similarly, when a compatibilist argues that we typically care about what is dependent on our executive motivational states (say, our choices), given that there are no special impairments in our capacity for choice, I would agree. When I deliberate, I am typically not thinking about philosophical accounts of my activity. But of course it would not follow that, upon reflection, I would reject the presupposition that my freedom must be the freedom to extend the actual past, given the laws of nature. Even when one is engaging in theorizing or philosophical reflection, one might not always put together different views and assemble a comprehensive picture. One might find some version of choice-dependence or the conditional analysis of freedom attractive; but the Consequence Argument can help to bring to bear implications of compatibilism that go against other beliefs one has (about the fixity of the past and natural laws). Upon reflection, one might find the views about the fixity of the past and natural laws even more basic than one's attraction to a suitably refined conditional analysis of freedom. (Of course, it might go the opposite way; that is, one might choose to reject the intuitive views about the fixity of the past and natural laws on behalf of a compatibilist account of freedom; note again that semicompatibilism in itself does not make a commitment one way or another here.)

Consider Gary Watson's contention that it is just obvious that a compatibilist must accept a notion of freedom that does not hold fixed the past and the natural laws; after all, the definition of causal determinism straightforwardly appears to imply this result. So why all the fuss about the Consequence Argument? Why suppose that it is somehow a revelation – a deeply problematic revelation – that the compatibilist must say that we (sometimes) have it in our power so to act that the past or the natural laws wouldn't be the same as they actually are?

Note first that it is typically not thought to be a defect in an argument that it is simple! And, of course, in fairness to Watson, he is not saying that this is a defect in the Consequence Argument; rather, he is making a comment on its dialectical role. Given the simplicity of the argument, Watson is wondering why its implications should be taken to be *interesting criticisms* of compatibilism – or perhaps new or surprising or revealing criticisms.

In reply, I would suggest that, as with most skeptical arguments, the Consequence Argument gets its grip by employing deeply appealing elements of commonsense. As I pointed out above, a proponent of the Consequence

Argument could be seen to be pointing out to a compatibilist that he must consider our intuitive views *comprehensively*; one does not always put together views that one nevertheless is inclined to accept. Sometimes an argument can bring out a troubling worry about a commitment by throwing it into relief in a new way, or even by associating it with a certain picture that renders its features salient.

It seems to me that our intuitive "picture" of the structure of possibility over time corresponds to Borges' idea of the future as a "garden of forking paths." That is, our intuitive conception or picture corresponds to a branching, treelike structure in which the various possible futures branch off a single line that can be traced back into the single actual past. When I deliberate, I assume that I have access to various possibilities that are connected to a single past; I do not assume that each possibility comes with its own past(s)! Intuitively, the future is a garden of forking paths; but each path branches off a single past (via a single present). To suppose that each future branch has its own past or set of pasts is to imagine a field overrun by weeds, and not an orderly garden. This picture is as unintuitive and unattractive as it is complex and inelegant. It seems to me that the idea that our freedom is the power to add to the given past, holding fixed the laws of nature, corresponds to important elements of our intuitive picture of agency or conceptual scheme, broadly construed.

Here is another (somewhat different) suggestion about what I take to be a genuinely puzzling aspect of the dialectical situation. Often compatibilists frame their discussions in terms of an attempt to give an "analysis" of the word "can," as it plays a certain signature role in discourse about free will and moral responsibility. This is a project that seeks to elucidate and regiment the meanings of our words. Similarly, sometimes the discussion is framed in terms of an attempt to articulate our inchoate "concept" of freedom (as it relates specifically to our concept of moral responsibility). These projects pertain to our language and our network of concepts. Here the "conditional analysis," perhaps suitably refined, is plausible – perhaps the notion of choice-dependence (where there are no impairments of the distinctive capacity to choose) – captures nicely what we *ordinarily mean* by the relevant "can," or perhaps it captures our *concept* of freedom (as it plays a specific role in our network of responsibility-concepts).

It is, however, well-established that it can be one thing to articulate a meaning or concept, and quite another to specify the nature or "real essence" of something. The meaning of the term, "water," and the ordinary concept, "water," presumably do not contain anything about "$H_20$." But arguably the nature or real essence of water is $H_20$. Similarly, the ordinary meaning of the term "can," and the ordinary concept of "freedom," may not contain anything

about the possibility of extending the actual past, holding the natural laws fixed; but arguably the nature or real essence of our freedom includes these features.

I have for many years been puzzled at how some philosophers find the Consequence Argument (in some form or another) absolutely and uncontroversially sound, whereas others dismiss it entirely. It is a weird feature of the discussions about free will and moral responsibility. One possible explanation of this puzzling phenomenon is that some philosophers are thoroughly focused on the issues about meanings and concepts, whereas others are attuned to the nature or real essence of freedom. (My first suggestion was that we sometimes do not put together all of our claims about a subject-matter; this second suggestion could be taken as the view that the compartmentalization comes from a difference in the kinds of claims that are at issue.)

The debates about whether the future is in fact a garden of forking paths, and whether we do in fact possess regulative control, are difficult and highly contentious. They have engaged serious and careful philosophers for millennia. The semicompatibilist, qua semicompatibilist, takes no stand here; that is, semicompatibilism is officially silent about whether (say) God's omniscience or causal determinism rules out regulative control. Rather, its distinctive claim is that causal determinism is compatible with the possession of a certain kind of control – guidance control – and moral responsibility, apart from whether causal determinism rules out regulative control. To someone who has absolutely no interest in a kind of freedom that involves the power to extend the past, holding fixed the laws of nature – someone who is not gripped at all by the ideas of the fixity of the past and natural laws – semicompatibilism will not be terribly interesting (in terms of his own thinking); it will simply be a different way of packaging views he already finds plausible. Of course, such an individual will *agree* with the basic thrust of semicompatibilism – that causal determinism is perfectly compatible with moral responsibility – and an invocation of the doctrine (in certain contexts) might well be *dialectically useful* for such a person. Indeed, semicompatibilism will be most helpful within a dialectical context in which some participants are indeed taken by the fundamental idea that our freedom is the power to add to the given past, holding fixed the laws of nature; it is important to see that a compatibilist about causal determinism and moral responsibility can *grant* a stringent interpretation of this idea. Given that millennia of debates have issued in a dialectical stalemate, semicompatibilism holds the promise of helping us to make real intellectual progress. If this is indeed old wine in new bottles, the possibility of progress in longstanding debates may be at least mildly intoxicating!

## 8 An Account of Guidance Control

I have sought to argue that moral responsibility does not require regulative control, but only guidance control, and further that it is plausible that guidance control is compatible with causal determinism. Although I have tried to render these claims attractive, I have not here attempted to give an account of guidance control. This is not the place to develop such an account in depth; rather, I shall simply sketch the outlines of this sort of account in order to give the reader the flavor of an "actual-sequence" theory of moral responsibility – an account that does not require that an agent *ever* have alternative possibilities with respect to choice, action, or even the formation of character.

On my approach to guidance control, there are two chief elements: the mechanism that issues in action must be the "agent's own," and it must be appropriately "reasons-responsive." Slightly more carefully, an agent exercises guidance control of his behavior insofar as it issues from his own, appropriately reasons-responsive mechanism. It is important to note that the invocation of the operation of "mechanism" here is not to reify anything – it merely refers to a process or a "way things go" along the path that leads to the behavior in question. Additionally, to fix on the way things go along this path is not in any way to de-emphasize or lose sight of the fact that the locus of control and moral responsibility is the *agent*; after all, the process in question is not only something that takes place (at least in part) in the agent, but it is "owned" by the agent.

Return to the Frankfurt-example presented above in which Jones votes for the Democrat on his own (for his own reasons and as a result of the normal human deliberative process). If Jones were about to choose to vote for the Republican, Black would intervene and cause (via direct electronic stimulation of the brain) Jones to choose to vote for the Democrat (and to go ahead and vote for the Democrat) anyway. The actual sequence and the alternative scenario involve intuitively *different kinds of mechanisms*: in the actual sequence, there is the normal operation of the human capacity for practical reasoning, whereas in the alternative scenario there is significant and direct electronic stimulation of the brain by the neurosurgeon. Even though it is difficult to provide a general account of mechanism individuation, it is (in my view) intuitively clear that different kinds of mechanisms operate in the actual and alternative sequences of the Frankfurt-cases. Further, it seems to me that what grounds the moral responsibility of the agent in such cases are features of the actual-sequence mechanism – properties of the path that actually leads to the behavior in question.

On my view, one relevant feature of the actual-sequence mechanism is that it must be in some appropriate way responsive to reasons. Note that it is

distinctive of the normal human capacity for deliberation that it is reasons-responsive. So even if the thorough electronic stimulation of Jones's brain of the sort applied in the alternative scenario (by Black) would issue in a choice to vote Republican, no matter what reasons there are for Jones to vote for the Democrat, Jones's actual-sequence mechanism is reasons-responsive. That is, in ascertaining reasons-responsiveness, one must hold fixed the actual-sequence mechanism, which, in the Frankfurt-case, is the normal human faculty of practical reasoning. So even if the agent (Jones) does not have genuine access to alternative possibilities (regulative control) in virtue of the existence of Black's set-up, he may well exhibit guidance control of his choice and voting behavior; after all, Black's set-up simply monitors the situation and does not play any role in Jones's choice and decision along the actual pathway.

It is a somewhat delicate matter to specify just the sort of reasons-responsiveness that must be present in order to have the sort of reasons-responsiveness that helps to ground moral responsibility. Elsewhere I have sought to give at least a sketch of an account of the relevant kind of reasons-responsiveness; here the details will have to be omitted, and "reasons-responsiveness" will have to remain vague. (See especially John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*.)

Now it seems to me that a suitably reasons-responsive mechanism could be "implanted" in a way that vitiates moral responsibility. Just as a neurosurgeon could directly stimulate the brain in such a way as to render the mechanism non-reasons-responsive, so the neurosurgeon could stimulate the brain in a more complex way that introduced a "new" pattern of reasons-responsiveness. This pattern is intuitively the scientist's, and not the agent's (Jones's). This extreme kind of scenario helps to show that a second crucial feature of the actual-sequence mechanism in moral responsibility is "ownership." In order to exhibit guidance control and be morally responsible, the actual-sequence mechanism must be the agent's own.

Here I have simply gestured toward the two main ingredients of the account of guidance control. In a sense, the basic intuitive ideas are more important than the details. Although I have not here offered detailed or more specific accounts of the relevant kind of reasons-responsiveness and also ownership, my contention is that they can in fact be developed in a way that is both plausible and arguably compatible with causal determinism.

My view then is that the "freedom-relevant" (as opposed to epistemic) condition of moral responsibility is guidance control. One can have guidance control of behavior without also having regulative control over it. Although the two kinds of control might typically go together, they can at least be analytically prized apart in contexts involving preemptive overdetermination (the Frankfurt-cases). An agent exercises guidance control insofar as his

behavior issues from his own, suitably reasons-responsive mechanism. I further contend that both mechanism-ownership and reasons-responsiveness are entirely compatible with causal determinism; thus, I contend that even if causal determination threatens regulative control, it is perfectly compatible with moral responsibility.

On this approach, what is important to moral responsibility is the actual history of the behavior under consideration. One looks at the properties of the actual-sequence mechanisms or processes that issue in behavior in assessing an agent's moral responsibility. Of course, these properties can be "modal" properties or sensitivities – such as reasons-responsiveness. But it is crucial that it is some feature of the actual path to the behavior – some (possibly modal) property or properties of *the way the behavior is actually generated*, rather than access to alternative pathways, that grounds moral responsibility.

As I have stated above, compatibilists traditionally contend that not all causally deterministic sequences equally threaten freedom and responsibility. Whereas I am inclined to accept the conclusion of the Consequence Argument and thus the view that all causally deterministic sequences equally rule out regulative control, I accept the view that not all causally deterministic sequences pose problems for guidance control (and thus moral responsibility). Now some compatibilists are content (perhaps reluctantly) to posit a mere list of "responsibility-undermining" factors, such as direct electronic stimulation or physical manipulation of the brain, certain kinds of hypnosis, brainwashing, aversive conditioning, subliminal advertising, drug interventions, irresistible impulses, unavoidable phobias, and so forth. These compatibilists contend that mere causal determination does not rule out moral responsibility, but they also want to say that the special circumstances specified in the (possibly-to-be-expanded) list *do* rule out responsibility. This approach is obviously not ideal or entirely satisfactory. In contrast, I have attempted to offer a general account of guidance control – an account that applies quite generally and also helps to explain why the special factors typically on the list do indeed rule out moral responsibility. Whereas some philosophers quite legitimately worry that it will not be easy to distinguish certain cases of problematic manipulation or causal influence from cases in which we are inclined to countenance moral responsibility, at least I have offered a general account that has some hope of sorting and distinguishing the range of possible cases in an illuminating, non-arbitrary fashion.

## 9 Conclusion: The Lure of Semicompatibilism

John Perry has told me that I need a new name for the position, and I agree that "semicompatibilism" is not very exciting. (My only consolation is that

the other names for positions in the Free Will debates are equally uninspiring; could you imagine going to the barricades for "hard incompatibilism"?) What is important, however, is not what's in the name, but what's in the doctrine. In this essay I have focused mainly on trying to explain the appeal of this form of compatibilism. That is, I have attempted to provide a *general* motivation for compatibilism, and also an explanation of the appeal of *this specific form* of compatibilism (as opposed to traditional compatibilism). The idea here has not been to develop detailed elaborations of the ideas or sustained defenses of the positions; rather, I have simply presented in sketchy form the attractions of the overall view (and the some of the difficulties faced by its competitors).

One of the main virtues of compatibilism is that our deepest and most basic views about our agency – our freedom and moral responsibility – are not held hostage to views in physics. A semicompatibilist would not have to revise these beliefs in light of a future discovery of the truth of causal determinism. Nor need he be prepared to revise his basic metaphysical views – such as that the past is fixed or the laws of nature are fixed or that powerlessness can be transmitted via the Principle of Transfer of Powerlessness – in light of such a discovery. A libertarian, it seems, must claim that he knows from his armchair that causal determinism is false; but how could we know in advance such an empirical thesis? These are significant virtues of semicompatibilism: a proponent of this doctrine need not purport to know *apriori* some (presumably) empirical thesis in physics, or be prepared to give up his basic views about our agency, or engage in unattractive "metaphysical flipflopping" (giving up some of one's basic metaphysical principles in light of some empirical truth in physics).

Semicompatibilism combines the best features of compatibilism and incompatibilism. Allen Wood has claimed that Kant believed that compatibilism and incompatibilism are consistent; this puzzling view can be defended insofar as here "compatibilism" and "incompatibilism" pertain to different "realms" or "perspectives" (phenomenal and noumenal). My doctrine of semicompatibilism does not show that compatibilism and incompatibilism are compatible (in the same realm), but it can accommodate the most compelling insights of the incompatibilist (as crystallized in the Consequence Argument) and also the basic appeal of compatibilism – that not all causally deterministic sequences equally rule out the sort of control that grounds moral responsibility. Thus, semicompatiblism allows us to track commonsense (suitably conceptualized in moral and legal theory) in making distinctions between those factors that operate in such a way as to undermine responsibility and those that do not. And a semicompatibilist need not give up the idea that sometimes individuals robustly deserve punishment for their behavior, whereas on other occasions they robustly deserve moral commendation and reward. That is, a

semicompatibilist need not etiolate or reconfigure the widespread and natural idea that individuals *morally deserve* to be treated harshly in certain circumstances, and kindly in others. We need not in any way damp down our revulsion at heinous deeds, or our admiration for human goodness and even heroism.

Semicompatibilism is both a conservative and radical doctrine. It is conservative in that it need not in any way call for revisions in the concept of moral responsibility or our actual responsibility practices, and it preserves the traditional idea that moral responsibility is associated with freedom or control. But it is radical in that it identifies guidance control, rather than regulative control, as the relevant sort of freedom. It thus departs significantly from traditional views in the *conditions* it posits for the application of the concept of moral responsibility (and thus the triggering of the responsibility-practices themselves).

In my view, we care deeply about being robustly free and morally responsible, and it is not straightforward to reconfigure our ideas or practices so that we eliminate residual retributive components in our attitudes to ourselves and others. Certainly, it is not easy to do so without a sense of loss. Semicompatibilism keeps a robust and traditional notion of moral responsibility. But the traditional picture is that we are morally responsible in virtue of *selecting* a path from among various paths that are genuinely open to us; in virtue of this selection, we *make a difference* of a certain sort to the world. The semicompatibilist denies that the value of our free agency – or the basis of our moral responsibility – is the power to make a difference. After all, it might turn out that we are mistaken in the natural and intuitive view that we have more than one genuinely available path into the future.

It may be that, just as there is a single line that connects the past to the present, there is only a single line into the future: a single metaphysically available path that extends into the future. In this case, what matters is how we proceed – how we walk down that path. There may be features that block access to alternative paths, but that play no role along the actual pathway. Thus, whatever it is that precludes access to alternative paths may not operate in such a way as to crowd out the features in virtue of which we are robustly morally responsible. When we walk down the path of life with courage, or resilience, or compassion, we might not (for all we know) make a certain sort of difference, but we *do* make a distinctive kind of statement. For the semicompatibilist, the basis of our moral responsibility is not selection in the Garden of Forking Paths, but self-expression in writing the narrative of our lives: it is not that we make a difference, but that we make a statement. In writing the stories of our lives, we connect the dots in a way that gives our lives a signature kind of meaning. Even if the name is unexciting, the idea is beautiful.

**Further Reading**

My contribution above relies heavily on my own previous work, as well as that of others. The essay is supposed to explain in a relatively informal way the main ideas in my approach to free will and moral responsibility and their motivations; I have thus not included footnotes. I hope the reader will understand that the nature of this book requires a less substantial scholarly apparatus than one might expect in another context. Below I make some very minimal suggestions for further reading, with an admittedly lamentable emphasis on my own more detailed elaboration of the ideas presented here.

I have attempted to present, defend, and elaborate semicompatibilism in John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control* (Oxford: Blackwell Publishers, 1994); John Martin Fischer and Mark Ravizza, S.J., *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press, 1998); and John Martin Fischer, *My Way: Essays on Moral Responsibility* (New York: Oxford University Press, 2006). *My Way* contains a more detailed discussion of my suggestion that the value of our free agency consists in a certain distinctive kind of self-expression. There is additional development of the idea that in acting freely we endow our lives with a signature sort of narrative value in: John Martin Fischer, "Free Will, Death, and Immortality: The Role of Narrative," *Philosophical Papers* (Special Issue: Meaning in Life), 34(3) (November 2005), 379–404. Some of the criticisms of source incompatibilism are based on material in John Martin Fischer, "The Cards That Are Dealt You," *Journal of Ethics*, 10 (2006), 107–29.

A landmark collection that explores and elaborates compatibilist themes is: Gary Watson, *Agency and Answerability* (Oxford: Clarendon Press, 2004). For important discussions of traditional compatibilist accounts of freedom and their difficulties, as well as presentations of the Consequence Argument, see Peter Van Inwagen, *An Essay on Free Will* (Oxford: Clarendon Press, 1983); and Carl Ginet, *On Action* (Cambridge, UK: Cambridge University Press, 1990).

Much of Harry Frankfurt's work on these subjects is collected in: Harry G. Frankfurt (ed.), *The Importance of What We Care About* (Cambridge, UK: Cambridge University Press, 1988). There are helpful discussions of various aspects of the Frankfurt-examples in David Widerker and Michael McKenna (eds.), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Aldershot, UK: Ashgate, 2003).

There are important discussions and defenses of source incompatibilism in Robert Kane, *The Significance of Free Will* (New York: Oxford University Press, 1996); and Derk Pereboom, *Living Without Free Will* (Cambridge:

Cambridge University Press, 2001). The latter book also contains insightful suggestions about the possibility of eliminating the retributive elements of our ordinary notion of moral responsibility. For additional worries about our ordinary and "robust" notion of responsibility, see Galen Strawson, *Freedom and Belief* (Oxford: Clarendon Press, 1986); and Saul Smilansky, *Free Will and Illusion* (Oxford: Clarendon Press, 2000).

There is an alternative development of an approach to moral responsibility that (apparently) does not require regulative control and that invokes a general capacity for reasons-responsiveness in R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1994).

# 3

# *Hard Incompatibilism*

*Derk Pereboom*

## 1 Outline of Hard Incompatibilism

Baruch Spinoza (1677/1985: 440–4, 483–4, 496–7) maintained that due to very general facts about the nature of the universe, we lack the sort of free will required for moral responsibility. I agree. More specifically, he argues that it is because of the truth of determinism that we lack this sort of free will; he is thus a *hard determinist*. By contrast, I am agnostic about the truth of determinism. I contend, like Spinoza, that we would not be morally responsible if determinism were true, but also that we would lack moral responsibility if indeterminism were true and the causes of our actions were exclusively states or events. If the causes of our actions were exclusively states or events, indeterministic causal histories of actions would be as threatening to moral responsibility as deterministic histories are. At the same time, I think that if we were undetermined agent-causes – if we as substances had the power to cause decisions without being causally determined to cause them – we might well then have the sort of free will required for moral responsibility. However, although agent causation has not been ruled out as a coherent possibility, the claim that we are agent-causes is not credible given our best physical theories. Thus we need to take seriously the prospect that we are not free in the sense required for moral responsibility. I call the resulting view *hard incompatibilism*. In addition, I argue that a conception of life without this sort of free will would not be devastating to morality or to our sense of meaning in life, and in certain respects it may even be beneficial.

Furthermore, I reject a type of incompatibilism according to which the availability of alternative possibilities is crucial to explaining moral responsibility, and accept instead a type of incompatibilism that ascribes the more

significant role to an action's causal history. I argue that an agent's moral responsibility for an action would be explained not by the existence of alternative possibilities available to her, but rather by the action's having a causal history of a sort that allows the agent to be the source of her action in a specific way. I thus opt for *source* as opposed to *leeway* incompatibilism. Agent-causal libertarianism is typically conceived as an incompatibilist position according to which an agent can be the source of her action in the way required for moral responsibility, and thus advocates of this view are typically source incompatibilists. But one might also be a source incompatibilist and seriously doubt that we have the sort of free will required for moral responsibility, and this is the stance I take.

The term "moral responsibility" is used in many ways, but there is one sense that has been at issue in the traditional philosophical debate about free will and determinism. My characterization of it is this: for an agent to be morally responsible for an action is for it to belong to her in such a way that she would deserve blame if she understood that it was morally wrong, and she would deserve credit or perhaps praise if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just because she has performed the action (given that she understands its moral status), and not by virtue of consequentialist considerations. This characterization leaves room for an agent's being morally responsible for an action even if she does not deserve blame, credit, or praise for it – if, for example, the action is morally indifferent.

There are other notions of moral responsibility. For example, an agent could be considered morally responsible if it is legitimate to expect her to respond to such questions as: "Why did you decide to do that? Do you think it was the right thing to do?" and to evaluate critically what her decisions and actions indicate about her moral character. The idea is that engaging in such interactions might well be reasonable in light of the way in which they contribute to our own and others' moral improvement (this notion is derived from Arthur Kuflik, in conversation; for a related conception, see Hilary Bok 1998: 151). But while this "legitimately called to moral improvement" notion may be a *bona fide* sense of moral responsibility, it is not the one at issue in the free will debate. For incompatibilists would not find our being morally responsible in this sense to be even *prima facie* incompatible with determinism. The notion that incompatibilists do claim to be at odds with determinism is rather the one defined in terms of basic desert.

In my view, the notion of moral responsibility at stake in the debate applies primarily to *decisions*; one might think of decision as the kind of action to which moral responsibility in this sense primarily applies. Intuitively, for an agent to be morally responsible for a decision in the "basic desert" sense

requires, crucially, that she have and be able to exercise a certain kind of control in its production. Having and being able to exercise this kind of control has traditionally been conceived as a kind of freedom of the will. It is sometimes assumed that free will is a matter of having alternative possibilities for decision, but this is not the only option. Instead, it might be thought to consist mainly in an agent's being the source of her decision in a particular way. I contend, then, that free will understood in this last way would provide the kind of control required for moral responsibility in the "basic desert" sense, but that it turns out that we do not have free will of this sort.

## 2 Alternative Possibilities

Why opt for a source as opposed to a leeway position? The intuition that an agent's moral responsibility for an action requires that she could have done otherwise has considerable force. This force is nicely expressed by what David Widerker calls the *W-defense*. About an agent (Jones) who breaks a promise, but could not have done otherwise, he writes:

> Still, since you, [Harry] Frankfurt, wish to hold him blameworthy for his decision to break his promise, tell me *what, in your opinion, should he have done instead*? Now, you cannot claim that he should not have decided to break the promise, since this was something that was not in his power to do. Hence, I do not see how you can hold Jones blameworthy for his decision to break the promise. (Widerker 2000: 191)

Despite the strong intuitions the W-defense captures, I think that examples of the kind devised by Frankfurt yield an effective challenge to the leeway position (Frankfurt 1969). In those examples, an agent considers performing some action, but an intervener is concerned that she will not actually come through. Thus if the agent were to show some sign that she will not or might not perform the action, the intervener would cause her to perform the action anyway. So an intervener, Black, might ensure that an agent, Jones, will perform an action, say, killing Smith, by implanting a device in her brain, which, upon detecting that she will or might not do so, would cause her to kill Smith nevertheless. In fact, however, Jones kills Smith on her own, without the intervention taking place. The intuition that Frankfurt aims to generate is that Jones could be morally responsible for killing Smith despite the fact that she could not have done otherwise.

However, some leeway incompatibilists have contended that a close examination of Frankfurt-style cases actually substantiates their position. For such cases involve some factor that the intervener's device is set up to detect that

could have but does not actually occur in the agent, such as *forming an intention* to do to otherwise. The possible occurrence of such a factor – a "flicker of freedom," to use John Fischer's term – is then proposed as the alternative possibility required for moral responsibility (Fischer 1994: 134–40).

Fischer argues, however, that one can construct Frankfurt-style examples in which the intervener's device detects some factor that occurs prior to the formation of the intention, and, more importantly, some factor that is not as intimately connected with the action itself. For instance, one might imagine that Jones will decide to kill Smith only if Jones blushes beforehand. Then her failure to blush (by a certain time) might be the alternative possibility that would trigger the intervention that would cause her to kill Smith. Supposing that Jones acts without intervention, we might well have the intuition that she is morally responsible for killing Smith, even though she could not have done otherwise than to kill him, and even though she could not even have formed an alternative intention. She could have failed to blush, but Fischer argues that such a flicker is of no use to the libertarian, since it is not sufficiently *robust* to have a role in grounding the agent's moral responsibility (Fischer 1994: 140–7).

I agree with Fischer that effective Frankfurt-style examples can be constructed in which any alternative possibility that remains is not robust. But what exactly is it for an alternative possibility to be robust? The key intuition underlying alternative possibility conditions is that if, for example, an agent is to be blameworthy for an action, it is crucial that she could have done something to avoid this blameworthiness. If the availability of an alternative possibility per se were in fact to play a role in explaining an agent's moral responsibility for an action, it would have to be robust at least in the sense that as a result of securing that alternative possibility instead, she would thereby have avoided the responsibility she has for the action she performed. It would be her securing of the alternative possibility that would explain why she would have avoided the responsibility she has. Failing to blush in the above scenario does not satisfy this criterion of robustness. If Jones had failed to blush, she would not thereby have avoided responsibility for killing Smith – it would not be the failure to blush itself that would explain why she would not be blameworthy. By typical incompatibilist intuitions, a robust sort of alternative possibility would at the very least involve the agent's willing to act in such a manner that would have precluded the action for which she is in fact morally responsible.

Robustness also has an epistemic component that needs to be made explicit. Imagine that the only way Jones could have voluntarily avoided killing Smith is by taking a sip from her coffee cup, and this is only because the coffee had been laced with a drug, ingestion of which would have replaced her desire to kill Smith with true love, as a result of which it would have been psychologi-

cally impossible for her to kill him. Suppose that she had no reason whatsoever to believe that taking the sip would preclude her killing Smith, because she had no reason to believe that the coffee was laced with the drug. In this situation, Jones could have voluntarily behaved in such a manner that would have precluded the action for which she was in fact blameworthy, as a result of which she would have avoided the moral responsibility she actually has (this reflects Michael Otsuka's proposed condition on blameworthiness, 1998; cf. Wyma 1997; McKenna 1997). But whether she could have voluntarily taken the sip from the coffee cup, without having any reason whatsoever to believe that it would have rendered her blameless in this way, is irrelevant *qua* alternative possibility to explaining why she is morally responsible for killing Smith. Despite the fact that Jones could have voluntarily taken a sip from her coffee cup, and doing so would have rendered her not morally responsible for killing Smith, this alternative possibility is nevertheless insufficiently robust to have an important role in grounding her moral responsibility.

This point suggests a more refined characterization of robustness. Here is a condition that I think is at least close to being correct:

> *Robustness*: For an alternative possibility to be relevant per se to explaining an agent's moral responsibility for an action, it must be that she could have willed something other than what she actually willed, such she correctly understood that by willing it she would thereby have been precluded from the moral responsibility she actually has for the action.

(This may not be quite right, since there may be examples of an agent who has a robust alternative possibility, where this alternative would preclude the responsibility she has for the option she selects, but due to some epistemic failing on her part, she does not believe that she has an alternative possibility that meets this specification. Dana Nelkin (in correspondence) suggests a case in which an agent mistakenly believes that the alternative possibility does not preclude the responsibility she has for the option she selects, but she does recognize significant morally salient differences between the two options. Here we may want to say that the agent has a robust alternative possibility partly because there are good reasons available to her for believing that she has an alternative in which her responsibility is different in the relevant way, even though she does not appreciate those reasons adequately. But this issue is complex, and I will leave it for another occasion.)

It might now seem that any alternative-possibilities condition on moral responsibility can be defeated by a Frankfurt-style example that employs a non-robust flicker of freedom. But this line of defense for such arguments has proven to be too quick. For it is challenged by an important objection to these sorts of arguments that was initially raised by Robert Kane and then

systematically developed by Widerker (Kane 1985: 51; 1996: 142–4, 191–2; Widerker 1995: 247–61; cf. Ginet 1996). The general form of the Kane/Widerker objection is this: for any Frankfurt-style scenario, if causal determinism is assumed to hold in that scenario, the libertarian will not have and cannot be expected to have the intuition that the agent is morally responsible. If, on the other hand, libertarian indeterminism is presupposed, an effective Frankfurt-style example is unavailable, for any such case will fall to a dilemma. In Frankfurt-style examples the actual situation will feature a prior sign that signals the fact that intervention is not necessary (such as the blush in Fischer's example). If in the proposed case the sign causally determined the action, or if it were associated with some factor that did, the intervener's predictive ability could be explained. However, then the libertarian would not have the intuition that the agent is morally responsible. If the relationship between the sign and the action were not causally deterministic in such ways, then the libertarian would object that the agent could have done otherwise despite the occurrence of the prior sign. Either way, some alternative possibilities condition on moral responsibility would emerge intact.

## 3 A Resilient Frankfurt-style Example

I have proposed a Frankfurt-style scenario that avoids the objections that have been raised for examples of this sort. Its distinguishing features are these: the cue for intervention – the flicker of freedom – must be a *necessary* rather than a sufficient condition, not for the action that the agent actually performs, but *for the agent's availing herself of any robust alternative possibility* (without the intervener's device in place), while the cue for intervention itself is not a robust alternative possibility, and the prior sign that the action will occur – the absence of the cue for intervention – clearly in no sense causally determines the action the agent actually performs. Here is the example:

> *Tax Evasion (2)*: Joe is considering claiming a tax deduction for the registration fee that he paid when he bought a house. He knows that claiming this deduction is illegal, but that he probably won't be caught, and that if he were, he could convincingly plead ignorance. Suppose he has a strong but not always overriding desire to advance his self-interest regardless of its cost to others and of even if it involves illegal activity. In addition, the only way that in this situation he could fail to choose to evade taxes is for moral reasons. He could not, for example, choose to evade taxes for no reason or simply on a whim. Moreover, it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to moral reasons. Joe can secure this level of attentiveness voluntarily. However, his attaining this level

> of attentiveness is not causally sufficient for his failing to choose to evade taxes. If he were to attain this level of attentiveness, he could, exercising his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place). However, to ensure that he will choose to evade taxes, a neuroscientist has, unbeknownst to Joe, implanted a device in his brain, which, were it to sense the requisite level of attentiveness, would electronically stimulate the right neural centers so as to inevitably result in his making this choice. As it happens, Joe does not attain this level of attentiveness to moral reasons, and he chooses to evade taxes on his own, while the device remains idle.

In this situation, Joe could be morally responsible for choosing to evade taxes despite the fact that he could not have chosen otherwise than to evade. (David Hunt also suggests this "necessary condition" strategy (2000), and develops a similar example (2005).)

The example does feature alternative possibilities that are available to the agent – his achieving higher levels of attentiveness to moral reasons. But these alternative possibilities are not robust. Note first that in ordinary circumstances, without the intervener's device in place, it is not the case that by achieving some higher level of attentiveness Joe would have avoided responsibility for choosing to evade taxes. For under these conditions achieving some higher level of attentiveness is compatible with his not refraining from making this decision, or even ever being seriously inclined so to refrain, and choosing to evade taxes instead. At this point one might object that given that the intervener's device is in place, by voluntarily achieving the specified higher level of attentiveness Joe would have voluntarily done something whereby he would have avoided the blameworthiness he actually incurs (Otsuka 1998). For had he voluntarily achieved the requisite level of attentiveness, the intervention would have taken place, whereupon he would not have been blameworthy for deciding to evade taxes. In reply, Joe does not understand, and, moreover, he has no reason to believe, that voluntarily achieving the requisite level of attentiveness would preclude him from responsibility for choosing to evade taxes, and hence this alternative possibility is not robust. True, were he voluntarily to achieve this attentiveness, the intervention would take place, and he would then not have been responsible for this choice. Nevertheless, Joe does not understand, and has no reason to believe, that the intervention would then take place, and that as a consequence he would be precluded from responsibility for this choice. In fact, one might imagine that he believes that achieving this level of attentiveness is compatible with his freely deciding to evade taxes anyway, and that he has no reason to believe otherwise. Nevertheless, Joe is morally responsible for deciding to evade taxes.

The key feature of the Kane/Widerker objection is that if the inevitability of the action given the prior sign is grounded in causal determinism, then the libertarian cannot be expected to agree that the agent is morally responsible for the action, but if we eliminate the causal determination then the agent has alternative possibilities after all. But here the inevitability of the action given the prior sign is not grounded in causal determinism, and at the same time no robust alternative possibilities are available to the agent (contrary to Stewart Goetz's criticism of the Tax Evasion example (2002: 146, n. 36). In this example, the relation between the prior sign and the action can be expressed as follows:

> If Joe fails to achieve a certain level of attentiveness to moral reasons, then, provided there is no intervention, he will decide to evade taxes.

The inevitability of Joe's decision is not grounded in causal determinism, since the absence of what would trigger the intervention at a particular time, that is, the absence of a certain level of attentiveness to moral reasons by a particular time, or a state indicated by this absence, does not, together with all the other actual facts about the situation, constitute a causally deterministic process that produces this decision. To see that this is so, imagine that the intervener and his device are removed from the situation. This is a legitimate move, since we've specified that the intervener and his device exert no actual causal influence on Joe's deciding to evade taxes, and so removing them will have no bearing on whether there is a causally deterministic process that produces his decision. Now notice that there is no relevant time at which refraining from deciding to evade taxes is not causally possible for him, since he can always attain the higher level of attentiveness, whereupon he can freely refrain from deciding to evade taxes – or else freely decide to evade taxes. Imagine that he does in fact decide to evade taxes, and he never achieves the specified level of attentiveness. Still, there is no causally deterministic process that issues in his deciding to evade taxes, for it is at no point causally determined that he will fail to achieve this level of attentiveness, and if he did achieve it, it would be causally open to him to refrain from deciding to evade taxes. In the last analysis, then, an agent can be morally responsible for an action even if no robust alternative possibilities is available to him, and the leeway position is in jeopardy.

What should we then say about the strong intuitions captured by Widerker's W-defense? Here I think that Michael McKenna has it right (2005a: 177). When Widerker asks of Joe, in view of the fact that he had no robust alternative possibility, "What would you have him do?," we should concede that there is no good answer. But against what we should admit to be this disturbing result, we should instead call attention to what Joe has actually done, and to the causal history by which his action came about.

## 4 Against Compatibilism

If this argument is successful, still both source incompatibilism and source compatibilism remain as live options. According to source compatibilism, compatibilism is true, and an agent's moral responsibility for an action is to be explained not by the existence of alternative possibilities available to her, but rather by the action's having a causal history of a sort that allows the agent to be the source of her action in a specific way. Fischer is an advocate of a view of this kind, and he is thus an opponent of source incompatibilism. While he noted the possibility of a source incompatibilist position early on (Fischer 1982), he maintains that "there is simply no good reason to suppose that causal determinism in itself (and apart from considerations pertaining to alternative possibilities) vitiates our moral responsibility" (1994: 159). Michael Della Rocca (1998: 102–3) contends that this is a claim for which Fischer has not argued, and in fact I believe he can be challenged on this point. True, one incompatibilist intuition that many seem naturally to have is that if we could not act otherwise than we actually do, then we could never refrain from the immoral actions we perform, and for this reason we would not be blameworthy for them. However, another very powerful and common intuition is that if all of our behavior were "in the cards" before we were born, in the sense that things happened before we came to exist that, by way of a deterministic causal process, inevitably resulted in our behavior, then we could not legitimately be judged blameworthy for our wrongdoing. By this intuition, if causal factors existed before a criminal was born that, by way of a deterministic process, inevitably issued in his act of murder, then he could not legitimately be judged blameworthy for his action. If all of our actions had this type of causal history, then we would seem to lack the kind of control over our actions that moral responsibility requires.

I do not believe that in the dialectic of the debate one should expect Fischer to be moved much by this incompatibilist intuition *alone* to abandon his compatibilism. Rather, the best type of challenge to the compatibilist at this point develops the claim that an action's being produced by a deterministic process that traces back to factors beyond the agent's control, even when she satisfies all the conditions on moral responsibility specified by the prominent compatibilist theories, presents in principle no less of a threat to moral responsibility than does deterministic manipulation. My "four-case argument" first of all develops examples that involve such manipulation, in which these compatibilist conditions on moral responsibility are met, and which elicit the intuition that the agent is not morally responsible. But further, it sets out three such cases, each progressively more like a fourth scenario, one that the compatibilist might envision to be realistic, in which the agent is causally determined to act in a natural way. The challenge for the

compatibilist is to point out a difference between this fourth scenario and one or more of the manipulation examples that shows why the agent might be morally responsible in the ordinary case but not in the manipulation examples. My suggestion is that non-responsibility generalizes from at least one of the manipulation cases to the ordinary one.

In each of the four cases I set out, Professor Plum decides to kill Ms White for the sake of some personal advantage, and succeeds in doing so. This act of murder conforms to the prominent compatibilist conditions, which are designed to be sufficient for an agent's moral responsibility when supplemented by some fairly uncontroversial additional necessary conditions. This act satisfies the various conditions proposed by David Hume and his followers: it is caused by desires that flow from his "durable and constant" character, since for him egoistic reasons typically weigh very heavily – much too heavily as judged from the moral point of view, while the desire on which he acts is nevertheless not irresistible for him, and in this sense he is not constrained to act (Hume 1739/1978: 319–412). It fits the condition proposed by Frankfurt: Plum's desire to murder White conforms to his second-order desires (i.e., desires to have some particular desire) in the sense that he wills to murder her and wants to will to do so, and he wills this act of murder because he wants to will to do so (Frankfurt 1971). The action meets the reasons-responsiveness condition proposed by Fischer and Ravizza: for instance, Plum's desires are modified by, and some of them arise from, his rational consideration of the reasons at issue, he is receptive to the relevant pattern of reasons, and if he knew that the bad consequences for himself that would result from killing White would be much more severe than they are actually likely to be, he would have refrained from killing her for this reason (Fischer and Ravizza 1998: 69–82). It also satisfies a condition proposed by Jay Wallace: Plum retains the general capacity to grasp, apply, and regulate his behavior by moral reasons. For example, when egoistic reasons that count against acting morally are relatively weak, he will typically regulate his behavior by moral reasons instead. These capacities even provide him with the ability to revise and develop his moral character over time (Wallace 1994: 51–83). Now, given that causal determinism is true, is it plausible that Plum is responsible for his action?

Each of the four cases that follow features different ways in which Plum's murder of White might be causally determined by factors beyond his control.

> Case 1: Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible given this history. These neuroscientists

> manipulate him to undertake the process of reasoning by which his desires are brought about and modified. They do this by pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum does not think and act contrary to character since his reasoning process is often manipulated to be rationally egoistic. His effective first-order desire to kill White conforms to his second-order desires. The process of deliberation from which his action results is reasons-responsive; in particular, this type of process would have resulted in his refraining from killing White in some situations in which the egoistic reasons were otherwise. Still, he is not exclusively rationally egoistic, since he typically regulates his behavior by moral reasons when the egoistic reasons are relatively weak – weaker than they are in the current situation. He is also not constrained in the sense that he does not act because of an irresistible desire – the neuroscientists do not provide him with a desire of this kind.

Plum's action satisfies all the compatibilist conditions we examined. But intuitively, he is not morally responsible because his action is causally determined by what the neuroscientists do, which is beyond his control. Consequently, it would seem that these compatibilist conditions are insufficient for moral responsibility, whether considered separately or in conjunction.

A compatibilist might resist this conclusion by arguing that although in Case 1 the process resulting in the action satisfies all of the prominent compatibilist conditions, yet Plum's relevant states are directly produced by the manipulators – he is locally manipulated – and this is the feature of the story that is responsibility-undermining. In reply, could a time lag between the manipulators' activity and the production of the states in the agent plausibly make a difference as to whether the agent is morally responsible? If all the manipulating activity occurred during one time interval and, after an appropriate length of time, these states were produced in him, could he only then be responsible? By my intuitions, such a time lag, all by itself, could make no difference to whether an agent is morally responsible for an action.

The strategy now requires a scenario more like the ordinary situation than Case 1. So here is Case 2 – which alone might also serve as a counterexample to the compatibilist conditions on morally responsible action:

> Case 2: Plum is like an ordinary human being, except that a team of neuroscientists has programmed him at the beginning of his life to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the consequence that in the circumstances in which he now finds himself, he is causally determined to undertake the reasons-responsive process of deliberation and to possess the set of first- and second-order desires that result in his killing White. Plum does have the general ability to regulate his behavior by moral reasons, but in his circumstances the egoistic reasons weigh heavily for

him, and as a result he is causally determined to murder White. Nevertheless, he does not act because of an irresistible desire.

Again, although Plum satisfies each of the compatibilist conditions, intuitively he is not morally responsible. Thus Case 2 also shows that our prominent compatibilist conditions, either separately or in conjunction, are not sufficient for moral responsibility. Furthermore, it would seem unprincipled to claim that here, by contrast with Case 1, Plum is morally responsible because the length of time between the programming and the action is great enough. Whether the programming takes place two seconds or thirty years before the action seems irrelevant to the question of moral responsibility. Causal determination by factors beyond Plum's control most plausibly explains his lack of moral responsibility in the first case, and I think we are forced to say that he is not morally responsible in the second case for the same reason.

Next consider a scenario more similar yet to the ordinary situation:

Case 3: Plum is an ordinary human being, except that he was determined by the rigorous training practices of his home and community so that he is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1 and 2). His training occurred when he was too young to have had the ability to prevent or alter the practices that determined his character. As a result, Plum is causally determined to undertake the reasons-responsive process of deliberation and to possess the first- and second-order desires that result in his killing White. He does have the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances the egoistic reasons are very powerful, and so the training practices of his upbringing, together with the background circumstances, deterministically result in his act of murder. Still, he does not act because of an irresistible desire.

If the compatibilist wants to claim that Plum is morally responsible in Case 3, he must point to a feature of these circumstances that would explain why he is morally responsible here but not in Case 2. However, it seems that there is no such feature. In each of these examples, Plum satisfies all of the prominent compatibilist conditions for morally responsible action, so a divergence in assessment of moral responsibility between these examples cannot be supported by a difference in whether these conditions are satisfied. Causal determination by factors beyond his control most plausibly explains his lack of moral responsibility in the second case, and we seem forced to conclude that Plum is not morally responsible in Case 3 for the same reason.

So it seems that Plum's exemption from responsibility in Cases 1 and 2 generalizes to the nearer-to-normal Case 3. Does it generalize all the way to the normal case?

> Case 4: Physicalist determinism is true, everything in the world is completely physical, and Plum is an ordinary human being, raised in normal circumstances, who is often but not exclusively rationally egoistic (just as egoistic as in Cases 1–3). Plum's act of killing White results from his undertaking the reasons-responsive process of deliberation, and he has the specified first- and second-order desires. He also possesses the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances the egoistic reasons weigh very heavily for him, and as a result he is causally determined to murder White. But it is not due to an irresistible desire that he kills her.

Given that we must deny moral responsibility to Plum in Case 3, could he be morally responsible in this more ordinary case? There would appear to be no differences between Case 3 and Case 4 that could support the claim that Plum is not morally responsible in Case 3 but is responsible in Case 4. One distinguishing feature of Case 4 is that the causal determination of Plum's crime is not brought about by other agents (Lycan 1997: 117–18). However, the claim that this is a relevant difference is implausible. Imagine a further case that is exactly the same as, say, Case 1 or Case 2, except that Plum's states are induced by a machine that is generated spontaneously, without intelligent design. Would he then be morally responsible? The compatibilist might agree that this sort of machine induction is responsibility-undermining as well, and then devise a condition that stipulates that agents are not responsible for actions manipulated by agents or machines. However, this move seems *ad hoc* – it appears motivated only by getting the desired compatibilist result. At this point the compatibilist might try to draw a line somewhere between agent manipulation and machine induction, but I don't see how this move could be developed in a plausible way.

The best explanation for the intuition that Plum is not morally responsible in the first three cases is that he lacks the control required for moral responsibility due to his action resulting from a deterministic causal process that traces back to factors beyond his control. Because Plum is also causally determined in this way in Case 4, we should conclude that here too Plum is not morally responsible for the same reason. More generally, if an action results from a deterministic causal process that traces back to factors beyond the agent's control, then he lacks the control required to be morally responsible for it.

By this argument, Plum's exemption from responsibility in Case 1 generalizes to his exemption from responsibility in Case 4. Notice that this argument is not a sorites – that is, its force does not depend on producing a series of cases, each of which is similar to its predecessor, and then arguing that since the first has some general feature, one must conclude that the last does as well because each of the successive pairs of cases are different only in some small degree of that type of general feature. A series of similar cases is indeed

important to the argument. But its strength derives from the fact that between each successive pair of cases there is no divergence at all in factors that could plausibly make a difference for moral responsibility, and that we are therefore forced to conclude that all four cases exhibit the same kind of responsibility-undermining feature.

A further variety of compatibilism, developed by P. F. Strawson, is also vulnerable to this sort of argument. He contends that the priority of practice – in this case the practice of holding people morally responsible – insulates attributions of moral responsibility from scientific or metaphysical challenges such as the one based on causal determinism (Strawson 1962). In my view, the best sort of argument against this position involves what Wallace calls a generalization strategy – arguing from generally accepted excuses or exemptions to the conclusion that causal determinism rules out moral responsibility (Wallace 1994). The excuses and exemptions that form the basis of this sort of argument would have to be generally accepted (but perhaps not uncontested), so that they are plausibly features internal to the practice of holding people morally responsible. The kinds of exemptions that I exploit in my argument are due to deterministic manipulation, and it is a feature of our practice that we exempt agents from moral responsibility when they are manipulated in this way, as in Cases 1 and 2. It is also a feature of our practice that if no morally relevant difference can be found between agents in two situations, then if one agent is legitimately exempted from moral responsibility, so is the other. No morally relevant difference can be found between agents in the manipulation examples and agents in ordinary deterministic situations such as Case 4. Thus it is the practice itself – in particular, key rules governing the practice – that renders moral responsibility vulnerable to causal determinism after all.

## 5 Two Objections to the Argument from Manipulation

Fischer has recently developed a challenge to this four-case manipulation argument. He contends that Plum *is* morally responsible in Cases 1 and 2, and that our intuition that he is not morally responsible stems from the correct sense that he is not *blameworthy*:

> In my view, further conditions need to be added to mere guidance control to get to blameworthiness; these conditions may have to do with the circumstances under which one's values, beliefs, desires, and dispositions were created and sustained, one's physical and economic status, and so forth. Professor Plum, it seems to me, is not blameworthy, even though he is morally responsible. That he is not blameworthy is a function of the circumstances of the

> creation of his values, character, desires, and so forth. But there is no reason to suppose that anything like such unusual circumstances obtain merely in virtue of the truth of causal determinism. Thus, I see no impediment to saying that Plum can be blameworthy for killing Mrs. White in Case 4. Note that there is no difference with respect to the minimal control conditions for moral responsibility in Cases 1 through 4 – the threshold is achieved in all the cases. But there are . . . wide disparities in the conditions for blameworthiness. (Fischer 2004: 158)

I agree that there are cases in which an agent is morally responsible without being blameworthy – when she is praiseworthy for having performed a morally exemplary action, or when she performs an action that is morally indifferent. But could Plum, who acts wrongly, and, we might suppose, understands that he does, be morally responsible without being blameworthy? In my view, an agent's being blameworthy for an action is in fact entailed by his being morally responsible for it in the sense at issue in the debate, together with his understanding that the action was in fact morally wrong. This is because for an agent to be morally responsible for an action in the sense at issue is for it to belong to him in such a way that he would deserve blame if he understood that it was morally wrong, and he would deserve credit or perhaps praise if he understood that it was morally exemplary, supposing that this desert is basic in the sense that the agent would deserve the blame or credit just because he has performed the action (given that she understands its moral status), and not by virtue of consequentialist considerations. Assuming this characterization, and Plum's understanding that killing White is morally wrong, he could not be morally responsible for committing this murder without also being blameworthy for it.

True, there are alternative senses of moral responsibility that allow Plum to be morally responsible and not blameworthy, such as the "legitimately called to moral improvement" notion, but they are not the ones at issue in this debate. If Plum is morally responsible in Case 1 and Case 2 in the "basic desert" sense that is at issue, then given his understanding that his action is morally wrong, it is entailed that he is blameworthy. An intuition that Plum can be morally responsible without being blameworthy might be explained by the possibility that he is responsible in some other sense while not being blameworthy. But it is not responsibility in these senses that incompatibilists have thought to be at odds with determinism.

McKenna challenges the four-case argument with a different kind of approach, one that involves foregrounding our intuitions about ordinary cases over those elicited by manipulation scenarios:

> The compatibilist's best strategy, it seems to me, is not to show how a suitably determined agent differs so very much from a globally manipulated agent. It

> is rather to show how similar they are. The compatibilist needs to make clear that once the manipulation is so qualified that all an agent's current time-slice compatibilist-friendly structures are properly installed through a process of manipulation, then the role of the manipulator begins to shrink into the background; we are simply left with a normal person who happened to be brought into existence in a very peculiar manner. Consider Derk Pereboom's use of global manipulation cases in his defense of incompatibilism. Pereboom wishes to start with manipulation cases, fix upon the hidden causes that seem to corrupt any appearance of responsibility, and then show how such cases are like standard cases of naturally occurring determination. Once the unseen causes of a naturally determined agent are revealed, Pereboom argues, then our reaction to the agent should be like our reaction to the discovery that a seemingly normally functioning agent has been globally manipulated. The compatibilist should meet Pereboom's challenge with two moves. First she should work in the other direction, *from* a (possible) naturally determined agent, *to* a globally manipulated one. Second, she should fix, not upon hidden causes, but upon the sorts of agential properties that typically serve as a basis for ascribing responsibility. Once it is established that actions issuing from a (possibly) naturally determined agent invite certain sorts of evaluations in terms of responsibility, one can then hold that actions issuing from an appropriately manipulated agent should be evaluated no differently. The nature of the hidden causes, it can thereby be argued, are not relevant to the sort of psychic structure on the basis of which an agent's responsibility is assessed. (McKenna, 2005b)

First a preliminary point. Part of the aim of the four-case argument is to foreground in our assessments of moral responsibility the causes of our actions that are not ordinarily evident, and in particular the (assumed) fact that they are deterministic. In our everyday moral judgments, we typically do not suppose that actions result from deterministic causal processes that trace back to factors beyond their agents' control. Our ordinary intuitions do not presuppose that causal determinism is true, and they could indeed presuppose that it is false. The incompatibilist's claim is that if we did assume determinism and internalize its implications, our judgments about moral responsibility might well be different from what they are. Spinoza remarks, "experience itself, no less than reason, teaches that men believe themselves free because they are conscious of their own actions, and ignorant of the causes by which they are determined" (Spinoza 1677/1985: 496). The sequence of manipulation cases is intended to generate the intuition that Plum is not morally responsible by making the deterministic nature of these causes salient. To claim that we should take our cue from examples in which the deterministic nature of the causes is not salient would beg the question against the incompatibilist, for that would amount to a refusal to engage his challenge.

Is it nevertheless possible to exert pressure on the four-case argument by stressing the fact that in everyday life compatibilist conditions count as

sufficient for moral responsibility? McKenna contends that one "should fix, not upon hidden causes, but upon the sorts of agential properties that typically serve as a basis for ascribing responsibility." But the incompatibilist could welcome fixing on those properties, for on his view they will often be necessary conditions for ascribing moral responsibility. However, this does not undermine the claim, which is made intuitive by the manipulation examples, that the absence of causal determination is also such a necessary condition. Notice that here the incompatibilist would recommend full disclosure: we should fix on *both* the hidden causes and the agential properties at issue. The suggestion that we should focus only agential properties would appear to be at a disadvantage in the dialectic of this debate.

McKenna's considered view is not that we should focus solely on the agential properties, but rather that in assessing the four-case argument, one could legitimately *draw greater attention* to them, and that this will elicit the intuition that Plum is responsible – certainly in Case 4, but even, for example, in Case 2 (McKenna, 2005b). At the same time, he allows that drawing greater attention to the hidden causes and their deterministic nature could generate the intuition that Plum is not morally responsible. But given that each of these two strategies is equally legitimate, the result will be a stalemate. In response, I advocate drawing *equal* attention to the sorts of agential properties that typically serve as a basis for ascribing responsibility, and to the hidden causes and their deterministic nature by means of the four cases, and then let the intuitions fall where they may. In fact, in my development of these cases the greater part of each description is devoted to setting out these agential properties. When I follow this recommendation, I retain the strong intuition that Plum in Case 4, and definitely in Case 2, is not morally responsible.

## 6 The Luck Objection to Event-causal Libertarianism

The position that remains is source incompatibilism. This view allows for an option that affirms the sort of free will required for moral responsibility – a kind of libertarianism – as well as one that denies it. I think that there are good reasons to be skeptical of the libertarian option.

There are two major versions of libertarianism, the event-causal and the agent-causal types. In event-causal libertarianism, actions are caused solely by way of states or events, and some type of indeterminacy in the production of actions by appropriate states or events is held to be a decisive requirement for moral responsibility (Kane 1996; Ekstrom 2000). According to agent-causal libertarianism, free will of the sort required for moral responsibility is accounted for by the existence of agents who possess a causal power to make

choices without being determined to do so. In this view, it is crucial that the kind of causation involved in an agent's making a free choice is not reducible to causation among states of the agent or events involving the agent, but is rather irreducibly an instance of a substance causing a choice not by way of states or events. The agent, fundamentally as a substance, has the causal power to make choices without being determined to do so (Chisholm 1976; O'Connor 2000; Clarke 2003).

Critics of libertarianism have contended that if actions are undetermined, agents cannot be morally responsible for them. The classical presentation of this objection is found in Hume's *Treatise of Human Nature*, and it has become known as the "luck "objection (Hume 1739/1978: 411–12; cf. Mele 2006). Let us consider the luck objection to the event-causal version of libertarianism. Intuitively, for an agent to be morally responsible for a decision, she must exercise a certain type and degree of control in making that decision. On an event-causal libertarian picture, the relevant causal conditions antecedent to a decision – agent-involving events, or, alternatively, states of the agent – would leave it open whether this decision will occur, and the agent has no further causal role in determining whether it does. With the causal role of these antecedent conditions already given, it remains open whether the decision will occur, and whether it will is not settled by anything about the agent – whether it be states or events in which the agent is involved, or the agent herself. So whether the decision will occur or not is, in this sense, a matter of luck. Accordingly, the agent lacks the control required for being morally responsible for the decision.

To illustrate, consider Kane's example of a businesswoman – let's call her Anne – who has the option of deciding to stop to assist an assault victim, whereupon she would be late for an important meeting, or deciding not to stop, which would allow her to make it to the meeting on time (Kane 1996). For simplicity, suppose the relevant antecedent conditions are, against stopping, *Anne's desiring at t not to annoy her boss*, and *Anne's believing at t that if she is late for the meeting her boss will be annoyed with her*; and for stopping, *Anne's desiring at t to help people in trouble*, and *Anne's belief that she can be effective in helping the assault victim*. Suppose the motivational force of each of these pairs of conditions is for her about the same. On an event-causal libertarian theory, with the causal role of these antecedent conditions already given, both Anne's deciding to stop and her deciding not to stop remain significantly probable outcomes. Suppose she in fact decides to stop. There is nothing else about Anne that can settle whether the decision to stop occurs, since in this view her role in producing a decision is exhausted by antecedent states or events in which she is involved. If nothing about Anne can settle whether the decision occurs, then she lacks the control required for moral responsibility for it. This might be called *the problem of the disappearing agent*. On an event-causal libertarian view

there is no provision that allows the agent to have control over whether the decision occurs or not (in the crucial sorts of cases), and for this reason she lacks the control required for moral responsibility for it.

Libertarians agree that an action's resulting from a deterministic sequence of causes that traces back to factors beyond the agent's control would rule out her moral responsibility for it. The deeper point of the luck objection is that if this sort of causal determination rules out moral responsibility, then it is no remedy simply to provide slack in the causal net by making the causal history of actions indeterministic. Such a move would fulfill one requirement for moral responsibility – the absence of causal determinism for decision and action – but it would not satisfy another – sufficiently enhanced control (Clarke 1997, 2003). In particular, it would not provide the capacity for an agent to be the source of her decisions and actions that, according to many incompatibilists, is ruled out by a deterministic framework.

## 7 Can Kane's Event-causal Libertarianism Evade the Luck Objection?

In Kane's variety of event-causal libertarianism, the paradigm sort of action for which an agent is morally responsible is one of moral or prudential struggle, in which there are reasons for and against performing the action in question. In his view, the production of such an action begins with the agent's character and motives, and proceeds through the agent's making an effort of will to act, which results in the choice for a particular action. The effort of will is a struggle to choose in one way given countervailing pressures, and it is explained by the agent's character and motives. In the case of a freely willed choice, this effort of will is *indeterminate*, and, consequently, the choice produced by the effort will be *undetermined*. Kane illuminates this last specification by drawing an analogy between an effort of will and a quantum event:

> Imagine an isolated particle moving toward a thin atomic barrier. Whether or not the particle will penetrate the barrier is undetermined. There is a probability that it will penetrate, but not certainty, because its position and momentum are not both determinate as it moves towards the barrier. Imagine that the choice (to overcome temptation) is like the penetration event. The choice one way or the other is *undetermined* because the process preceding it and potentially terminating in it (i.e. the effort of will to overcome temptation) is *indeterminate*. (Kane 1996: 128)

The effort of will is indeterminate in the sense that its causal potential does not become determinate until the choice occurs. Prior to this pivotal

interaction, there are various ways in which this causal potential can be resolved, and thus when it is resolved, the choice that ensues will be undetermined. Kane cautions against construing his view in such a way that the indeterminacy occurs after the effort is made: "One must think of the effort and the indeterminism as fused; the effort is indeterminate and the indeterminism is a property of the effort, not something that occurs before or after the effort." He contends that if an agent is morally responsible for a choice, either it must be free in this sense or there must be some such free choice that is its sufficient ground, cause, or explanation (1996: 35). In addition, Kane elaborates his position by an account as to how the particle analogy for free choice might actually work in the functioning of the brain's neural networks (1996: 128–30).

In response to the luck objection, Kane cites several commentators who have pointed out that an agent can in fact be responsible for an event that is indeterministic (Austin 1966; Foot 1957; Anscombe 1971). He provides a convincing example: it may not be causally determined that the radioactive material the employee places in the executive's desk will give him cancer, but this absence of causal determination is consistent with the employee's being morally responsible for the executive's developing cancer if he in fact does (1996: 55). Kane is clearly right about this. Still, as Galen Strawson contends, if the indeterminism is located at a different point in the production of an action than it is in this example, moral responsibility is more clearly threatened (1986: 8). Imagine that the indeterminism is located not in the consequences of the decision, as in the example, but in the decision to act. If *the agent's deciding to place the radioactive material in the desk* is an indeterministic event, then perhaps it is no longer clear that the employee is morally responsible.

This suggestion must be developed with care. One might imagine versions of event-causal libertarianism allowing agents to be responsible for decisions characterized by two distinct types of causal histories. In the first sort,

(a) events of type
    (E) *an agent's being in circumstances C,* where C includes the agent's character, desires, beliefs, and external circumstances
cause events of type:
    (F) *the agent's deciding to do A.*
but E's do not deterministically cause F's.

In the second sort,

(b) events of type (F) occur without being caused at all.

Events with either sort of causal history will be indeterministic. On a first pass at a luck objection to Kane's view, agents will not be morally responsible for decisions with either sort of causal history because they are *not sufficiently within the agent's control*.

Kane has developed two lines of defense against this version of the luck objection. First, he argues that decisions can be undetermined and yet have many features indicative of agent control and moral responsibility. Undetermined decisions could still be made voluntarily, intentionally, knowingly, on purpose, and as a result of the agent's efforts. They might still be made for reasons; agents might make them for reasons, rather than by mistake, accident, or chance; and agents may want to make them for these reasons rather than any others. Agents might not be coerced or compelled in making undetermined decisions, and in making them they might not be controlled by other agents or circumstances (Kane 1996: 179; 1999: 237–9). The absence of causal determination is thus consistent with significant control in action, and, Kane contends, with control sufficient for moral responsibility.

The fact that indeterministic action can have these characteristics does show that it allows for significant control in action. However, compatibilists appeal to the very same characteristics to defend against the objection that causal determinism precludes control sufficient for responsibility. Since incompatibilists deny that causally determined action can feature control sufficient for responsibility, one might question whether Kane's first strategy can, in the last analysis, deliver the control incompatibilists want.

Kane's second strategy for defending event-causal libertarianism appeals to the phenomenology – the introspectively experienced character – of choice and action. He contends that if initiation of an undetermined action were experienced as an uncaused, involuntary event, not resulting from one's effort of will, then we would have strong reason to believe that no genuine choice is involved, and that the agent is not morally responsible for the action. But if the initiation of an undetermined action were experienced as voluntary and as resulting from one's own effort of will, then the agent's moral responsibility would not be undercut. The example of Anne and the assault victim features an inner struggle between Anne's moral conscience, which urges her to stop and help the victim, and her career ambitions, which tell her she cannot miss her meeting. When the struggle is resolved in favor of the decision to stop and help the victim, Kane supposes that the effort of will from which the decision results is indeterminate, and that consequently the decision in undetermined. He then remarks:

> Now indeterminism may in some instances undermine choice . . . We imagined that Jane had reached a point in her deliberation at which she favored

> vacationing in Hawaii when, owing to a quantum jump in her brain, she found herself intending to vacation in Colorado. The case was odd because she did not have the sense of voluntarily doing anything . . . she would be reluctant – and we would be reluctant – to say she *chose* anything in such a case . . . So indeterminism can sometimes undermine choice. But there is no legitimate reason to generalize from cases like Jane's and say it must always do so. Consider [Anne] the businesswoman by contrast. Her experience, unlike Jane's, is of consciously and voluntarily choosing to follow her moral conscience and to return to help the victim, thereby resolving a preceding uncertainty in her mind. Also, in the businesswoman's case, unlike Jane's, the indeterminate process discovered by the neuroscientists immediately preceding the choice was experienced by her as her own effort of will, not merely as a random occurrence in her brain that happened to influence the outcome. Given these circumstances, it would be hasty, to say the least, to lump the two cases together and draw conclusions about the businesswoman's case from Jane's . . . Why would the businesswoman conclude that she did not really choose in such circumstances (rather than that her choice was undetermined) just because, *under very different circumstances,* Jane did not really choose? (1996: 182–3)

In Kane's view, then, the phenomenology of decision-making process is decisive or at least counts heavily for settling whether an agent is morally responsible for an undetermined action.

However, a compatibilist could as easily appeal to this same sort of phenomenological consideration in response to the objection that agents cannot be responsible for causally determined actions, and this should be of concern to the incompatibilist. A compatibilist might argue that if an agent experienced her causally determined decision as resulting from an outside determining force, she would have good reason to believe that she was not making a choice for which she was morally responsible. If, by contrast, a causally determined decision were experienced as voluntary and resulting from the agent's effort of will, she would have a strong reason to believe she was morally responsible for it. But incompatibilists would generally reject this compatibilist defense for the reason that a metaphysical fact about the causal history of the action – that the decision is causally determined – rules out the agent's moral responsibility, regardless of its phenomenology. It would be implausible to claim that the phenomenology can carry more weight when the threat to moral responsibility is indeterministic rather than deterministic.

Because they are available to compatibilists, Kane's two strategies give rise to a more refined version of our objection to event-causal libertarianism. I will grant, for purposes of argument, that this position allows for *as much* control as does compatibilism (Clarke, 1995, 2000, 2003). However,

following Randolph Clarke's suggestion (1997, 2003), the concern is that if decisions were indeterministic events, then agents would have *no more control* over their actions than they would if determinism were true, and such control is insufficient for responsibility.

This concern might not threaten a position such as Kane's if the only reason why causal determinism undermines moral responsibility is that it rules out alternative possibilities. For then Kane might argue that only alternative possibilities need to be added to the sort of control that a deterministic account of action can provide, and this his view can deliver. However, if a Frankfurt-style argument is successful, and I have contended that it is, then it can't be that the only reason causal determinism undermines responsibility is that it rules out alternative possibilities. My alternative proposal is that if an agent's decision is causally determined by factors beyond her control, she will not be the source of her decisions in a way that allows for moral responsibility. The inability of event-causal libertarianism to provide this type of sourcehood is strongly suggested by the luck objection, and in particular by the problem of the disappearing agent.

We might amplify this more refined objection by turning to Kane's UR (for "ultimate responsibility"), which specifies his key conditions for moral responsibility. UR has two elements. The first, in essence (Kane's formulations of these components are more precise (1996: 35)), is that to be ultimately responsible for an event, an agent must have voluntarily been able to do otherwise. The second is that to be ultimately responsible for an event, an agent must be responsible for any sufficient ground or cause or explanation of the event. If actions are undetermined events, then the first component of UR might be satisfied, and agents could have the required leeway for alternative actions. For Kane the second component is grounded in a more fundamental requirement about the origination of action:

> (Q_) If the action did have such a sufficient reason for which the agent was not responsible, then the action, or the agent's will to perform it, would have its source in something that the agent played no role in producing . . . ultimately responsible agents must not only be the sources of their actions, but also of the *will* to perform the actions. (1996: 73)

I think (Q_) expresses the deepest and most plausible source-incompatibilist intuition. At the same time, however, it also yields a threat to Kane's position. First, (Q_) has the consequence that agents cannot be responsible for decisions that are undetermined because they are not produced by anything at all, for then agents quite obviously cannot be the source of the will to perform them. This consideration subverts event-causal libertarianisms according to which agents can be responsible for decisions of type (b) above, in which events of

type (F), viz. *the agent's deciding to do A,* occur without being caused at all. We shall have to see whether it also undermines a view like Kane's.

Between decisions that are undetermined because they are not produced by anything at all, and those that are causally determined by factors beyond the agent's control, lies a range of decisions for which factors beyond the agent's control contribute to their production but do not determine them, while there is nothing that supplements the causal contribution of these factors to produce them. By analogy, according to the standard interpretation of quantum mechanics, antecedent events causally influence which quantum event will occur from among a range of possibilities by fixing the probabilities governing this range, but these antecedent events do not causally determine which of these possible quantum events will occur. So similarly, antecedent events might influence which decision an agent will make without determining any particular decision.

However, given the source-incompatibilist intuition, agents cannot be morally responsible for decisions if they meet these specifications. If there are factors beyond the agent's control that influence a decision's production without causally determining it, while there is nothing that supplements the contribution of these factors to produce the decision, then its production features only a combination of the first two types of responsibility-undermining factors. We have already seen that by the source-incompatibilist intuition, agents are not morally responsible for decisions causally determined by factors beyond their control. However, if these factors, rather than determining a single decision, simply leave open more than one possibility, and the agent plays no further role in determining which possibility is realized, then we have no more reason to believe that she is morally responsible than in the deterministic case.

Let us designate those events for which factors beyond the agent's control determine their occurrence *alien-deterministic events*, and those that are not caused by anything at all *truly random events*. And we might call the events in the range between these two extremes – those for which factors beyond the agent's control causally influence their production but do not causally determine them, while there is nothing that supplements the contribution of these factors to produce them – *partially random events*. By the source-incompatibilist intuition, an agent cannot be morally responsible for a decision if it is an event that lies anywhere on this continuum, because the agent does not have a suitable role in its production – the agent will not be the source of such a decision in a sense sufficient for moral responsibility. But it seems that for Kane free decisions are in fact partially random events.

At this point one might argue that there is an additional resource available for Kane's account of morally responsible decision. It could be suggested that in his conception, the character and motives that explain an effort of will

need not be factors beyond the agent's control, since they could be produced partly as a result of the agent's free choice. Consequently, it need not be that the effort, and thus the choice, is produced solely by factors beyond the agent's control and not by the free choice of the agent.

But this move faces a challenge. To simplify, suppose that it is character alone and not motives in addition that explains the effort of will. Imagine first that the character that explains the effort is not a product of the agent's free choices, but, rather, that there are factors beyond her control that determine this character, or nothing produces it, or factors beyond her control contribute to the production of the character without determining it and nothing supplements their contribution to produce it. Then the agent will not be responsible for her character. In addition, neither can she be responsible for the effort that is explained by the character, whether this explanation is deterministic or indeterministic. If the explanation is deterministic, then there will be factors beyond the agent's control that determine the effort, and the agent will thereby lack moral responsibility for the effort. If the explanation is indeterministic, given that the agent's free choice plays no role in producing the character, and nothing besides the character explains the effort, there will be factors beyond the agent's control that make a causal contribution to the production of this effort without determining it, while nothing supplements the contribution of these factors to produce the effort. Then again, the agent will not be morally responsible for the effort.

However, prospects for moral responsibility for the effort of will are not enhanced if the agent's character is in part a result of her free choices. For consider the first free choice an agent ever makes. By the above argument, she cannot be responsible for it, since she cannot be responsible for the effort of will from which it results. But then she cannot be responsible for the second choice either, whether or not the first choice was character-forming. If the first choice was not character-forming, then the character that explains the effort of will that results in the second choice is not produced by her free choice, and then by the above argument, she cannot be morally responsible for it. Suppose, alternatively, that the first choice was character-forming. Because the agent cannot be responsible for the first choice, she also cannot be responsible for the resulting character formation. And then, by the above argument, she cannot be responsible for the second choice either. Since this type of reasoning can be repeated for all subsequent choices, an agent that meets Kane's specifications will never be morally responsible for an effort of will.

Since such an agent will never be morally responsible for an effort of will, she will also not be responsible for the resulting choices. The reason is that on Kane's account, there is nothing that supplements the contribution of the effort of will to produce the choice. In fact, all free choices will be partially

random events, for there will be factors beyond the agent's control, such as her initial character, that causally influence which choice is produced without causally determining it, while nothing supplements the contribution of these factors in the production of the choice. It might be claimed that if decisions were underlain by complex, perhaps chaotic arrangements of such events, the enhanced control would emerge (Kane 1996: 128–30). However, if decisions themselves are at best partially random events, agents will not have the enhanced control required for moral responsibility, despite the proposed complexity of the decision's underlying structure.

## 8 The Luck Objection to Agent-causal Libertarianism

What needs to be added to the event-causal libertarian story is involvement of the agent in the production of decisions that would enhance her control so as to make it sufficient for moral responsibility. This enhanced control would remedy the problem of the disappearing agent that is highlighted by the luck objection against event-causal libertarianism. The agent-causal libertarian's proposal is to reintroduce the agent as a cause, this time not merely as involved in events, but rather fundamentally as a substance. The agent-causal libertarian thus appeals to the controversial notion of substance causation (for the reasons why it is controversial, see Clarke 2003: 196–210). If the agent were reintroduced merely as involved in events or states, the arguments already raised against the adequacy of event-causal libertarianism could be reiterated with undiminished effect. What the agent-causal libertarian posits is an agent who possesses a causal power, fundamentally as a substance, to cause a decision without being causally determined to do so. The proposal is that the control absent on the event-causal libertarian view – the sort of control sufficient for moral responsibility – is supplied by the agent by virtue of having this causal power.

One should note here that an adequate conception of this causal power would have to include additional elements. For example, provision needs to be made that it can be exercised rationally. An option here is to build this provision into the causal power itself – it might be characterized as a causal power, fundamentally as a substance, to cause a decision upon consideration of reasons, and on the basis of certain reasons, without being causally determined to do so.

Agent-causal libertarianism, I contend, does not fall to the luck objection. Recently, Carl Ginet, Alfred Mele, and Ishtiyaque Haji have each disputed this contention, on the grounds that the luck objection has as much force against the claim that agent-causal libertarianism allows for control sufficient for moral responsibility as it does against the view that event-causal

libertarianism allows for this sort of control (Ginet 1997; Haji 2004; Mele 2005, 2006). In response, it is first of all indisputable that when an agent A agent-causes decision D at time t, then an event of the following type occurs:

G: A's causing D at t.

Mele and Haji point out that on the agent-causal picture, given exactly the same conditions antecedent to t as those that precede A's agent-causing D, G might not have occurred. Hence, they argue, the fact that G did occur is a matter of luck. As a result, the agent does not have control in making decision D that is sufficient for moral responsibility.

The agent-causal libertarian will certainly admit that given these antecedent conditions, G might either have occurred or not. But the core issue is whether the agent nevertheless can have the crucial role in making decision D that she cannot have on the event-causal libertarian's view. And it has not been ruled out that she can. What the agent-cause does *most fundamentally* is to cause a decision. At this point, one should note that it is a *logical consequence* of the agent's causing a decision that an event of type G occurs. It follows logically from the fact that Anne, now equipped with the agent-causal power, causes the decision to stop that the event *Anne's causing the decision to stop* occurs. But it is *by* agent-causing a decision that the agent brings about the event of type G – as a logical consequence of her causing the decision. What thus explains the occurrence of the event of type G – indeed, given that the antecedent events are in place – is Anne, as a substance, causing the decision. This account differs from the event-causal libertarian's scenario, where, given the role of the antecedent events, the agent plays no further part in settling whether the decision occurs (cf. Clarke 2003: 105; Pereboom 2001: 48, 2004).

## 9 Agent Causation and Physics

Can agent-causal libertarianism be reconciled with what we would expect given our best physical theories? If this theory is true, then when we make free decisions, we cause these decisions without being causally determined to do so. As agent-causes that fit this description, we at some point would affect the physical world external the agent-cause, perhaps most directly in the brain. If the physical world were generally governed by deterministic laws, it seems that at the point of interaction we would then encounter divergences from these laws. One might object that it is possible that every indeterministic choice ever made just happens to dovetail with what could in principle be predicted on the basis of the deterministic laws, so that there are actually no

divergences from these laws. But this proposal involves coincidences too wild to be believed. For this reason, agent-causal libertarianism is not credibly reconciled with the physical world's being governed by deterministic laws.

On the standard interpretation of quantum mechanics, however, the physical world is not in fact deterministic, but is rather governed by probabilistic statistical laws. Clarke has defended the claim agent-causal libertarianism can be reconciled with physical laws of this sort (Clarke 1993: 193; 2003: 181). However, wild coincidences would also arise on this suggestion. Consider the class of possible actions each of which has a physical component whose antecedent probability of occurring is approximately 0.32. It would not violate the statistical laws in the sense of being logically incompatible with them if, for a large number of instances, the physical components in this class were not actually realized close to 32 percent of the time. Rather, the force of the statistical law is that for a large number of instances it is correct to *expect* physical components in this class to be realized close to 32 percent of the time. Are free choices on the agent-causal libertarian model compatible with what the statistical law leads us to expect about them? If they were, then for a large enough number of instances the possible actions in our class would almost certainly be freely chosen close to 32 percent of the time. But if the occurrence of these physical components were settled by the choices of agent-causes, then their actually being chosen close to 32 percent of the time would amount to a coincidence no less wild than the coincidence of possible actions whose physical components have an antecedent probability of about 0.99 being chosen, over a large enough number of instances, close to 99 percent of the time. The proposal that agent-caused free choices do not diverge from what the statistical laws predict for the physical components of our actions would run so sharply counter to what we would expect as to make it incredible.

Clarke contends that "If there can be substance causation at all, then it seems that there can be substance causation the propensities of the exercise of which conform with complete nondeterministic microlevel causal laws" (2003: 181). I think that Clarke is right about this. It is *possibl*e that the agent causation accord with probabilistic microlevel laws, for it might just happen that in the long run the exercise of agent-causal powers conforms to the probabilities that the indeterministic microlevel laws would assign in the absence of agent causation.

But should one *expect* this conformity? Timothy O'Connor argues that if the antecedent events are conceived as shaping the agent-causal power, then it is reasonable to expect the actions of agent-causes to conform to the probabilities conferred by these antecedent events:

> Imagine that some conscious reasons-guided systems "magnify" microphysical indeterminacies in such a way that several significantly different outcomes are

> physically possible. Then further suppose that agent-causal power emerges when conscious reasons-guided systems achieve a requisite threshold of complexity. Such a power might be shaped by states (such as the agent's reasons for acting) that embody the magnified quantum indeterminacies, so that agent-causal actions would be expected to reflect the physical probabilities in the long run. (O'Connor 2003: 309; cf. Clarke 2003: 181)

However, to answer the luck objection, the causal power exercised by the agent must be of a different sort from that of the events that shape the agent-causal power, and on the occasion of a free decision, the exercise of these causal powers must be distinct from the exercise of the causal powers of the events. For the luck objection shows that causal powers of the events are not the sort that can provide the control needed for moral responsibility. The agent as substance, by virtue of her agent-causal power, is meant to provide this sort of control. Given this requirement, we would expect the decisions of the agent-cause to diverge, in the long run, from the frequency of choices that would be extremely likely on the basis of the events alone. If we nevertheless found conformity, we would have very good reason to believe that the agent-causal power was not of a different sort from the causal powers of the events after all, and that on the occasion of particular decisions, the exercise of these causal powers was not distinct from the exercise of the causal powers of the events. Accordingly, the shaping that O'Connor has in mind cannot be so radical as to undercut the independence of the agent-causal power from the causal powers of the events, and if it is not, then we would expect the divergences at issue.

At this point, the libertarian might propose that there actually do exist divergences from the probabilities that we would expect without the presence of agent-causes, and that these divergences are to be found in the brain. The problem for this proposal, however, is that we have no evidence that such divergences occur. This difficulty, all by itself, provides a strong reason to reject this approach.

It is sometimes claimed that we have significant phenomenological evidence for the broader thesis that we have libertarian free will. Perhaps, then, if we could have libertarian free will only if we were agent-causes, and if we were agent-causes there would exist the divergences at issue, then our phenomenological evidence counts in favor of the existence of these divergences. However, the Spinozan response to this claim, that we believe our decisions are free only because we are ignorant of their causes, has not been successfully countered. The lesson to draw from Spinoza is that the phenomenology apt to generate a belief that we have libertarian free will would be just the same if decisions were instead causally determined and we were ignorant of enough of their causes. For this reason, the phenomenological evidence for our having libertarian free will is not especially impressive. This consideration counts

strongly against the proposal that such evidence gives us reason to believe that the divergences in question exist.

On the other hand, nothing we've said conclusively rules out the claim that because we are agent-causes, there exist such divergences. We do not have a complete understanding of the human neural system, and it may turn out that some human neural structures differ significantly from anything else in nature we understand, and that they serve to ground agent causation. This approach may be the best one for libertarians to pursue. But at this point we have no evidence that it will turn out to be correct.

Thus all versions of libertarianism face serious difficulties. Earlier, we raised problems for the leeway positions, and for compatibilism. Hence the only position in the free will debate for which problems have not been raised is the version of source incompatibilism that denies we have the sort of free will required for moral responsibility. This is a variety of hard incompatibilism. The concern for this view is not, I think, that there is significant empirical evidence that it is false, or that there is a good argument that it is somehow incoherent, and false for that reason. Rather, the questions it faces are practical: What would life be like if we believed it was true? Is this a sort of life that we can tolerate?

## 10 Hard Incompatibilism and Wrongdoing

Accepting hard incompatibilism requires denying our ordinary view of ourselves as blameworthy for immoral actions and praiseworthy for actions that are morally exemplary. At this point one might object that giving up our belief in moral responsibility would have very harmful consequences, perhaps so harmful that thinking and acting as if hard incompatibilism is true is not a feasible option. Thus even if the claim that we are morally responsible turns out to be false, there may yet be weighty practical reasons for continuing to treat people as if they were. For example, perhaps treating wrongdoers as blameworthy is often required for effective moral education and improvement. If we resolved never to treat people as blameworthy, one might fear that we would be left with insufficient leverage to reform immoral behavior.

Still, this option would have us treat people as blameworthy – by, for example, expressing indignation towards them – when they do not deserve it, which would seem prima facie morally wrong. As Bruce Waller argues, if people are not morally responsible for immoral behavior, treating them as if they were would be unfair (Waller 1990: 130–5). However, it is possible to achieve moral reform by methods that would not be threatened by this sort of unfairness, and in ordinary situations such practices could arguably be as

successful as those that presuppose moral responsibility. Instead of treating people as if they deserve blame, the hard incompatibilist can turn to moral admonition and encouragement, which presuppose only that the offender has done wrong. These methods can effectively communicate a sense of right and wrong and result in genuine reform.

However, does hard incompatibilism have resources adequate for addressing criminal behavior? Here it would appear to be at a disadvantage, and, if so, practical considerations might force us nevertheless to treat criminals as if they were morally responsible. If hard incompatibilism is true, a retributivist justification for criminal punishment is unavailable, for it assumes that the criminal deserves pain or deprivation just for committing the crime, while hard incompatibilism denies this claim. And retributivism is one of the most naturally compelling ways for justifying criminal punishment.

By contrast, a theory that justifies criminal punishment on the ground that punishment educates criminals morally is not challenged by hard incompatibilism specifically. So one might propose that the hard incompatibilist could appeal to a view of this kind. But we lack significant empirical evidence that punishment of criminals brings about moral education, and without such evidence, it would be wrong to punish them for the sake of achieving this goal. In general, it is wrong to harm someone in order to realize some good in the absence of significant evidence that the harm will produce the good. Moreover, even if we had good evidence that punishment was effective in morally educating criminals, we should prefer non-punitive methods for achieving this result, if they are available – whether or not criminals are morally responsible.

Deterrence theories claim that criminal punishment is justified on the ground that it deters future crime. The two most prominent deterrence theories, the utilitarian version and the one that grounds the right to punish on the right to self-defense, are not undercut by hard incompatibilism per se. However, they are questionable on other grounds. The utilitarian theory, which claims that punishment is justified because it maximizes utility, faces well-known challenges. It would seem at times to demand punishing the innocent when doing so would maximize utility; in some circumstances it would appear to prescribe punishment that is unduly severe; and it would authorize harming people merely as means to the well-being, in this case the safety, of others. The type of deterrence theory that founds the right to punish on the right of individuals to defend themselves against immediate threats (Farrell 1985: 38–60) is also objectionable. For when a criminal is sentenced he is typically not an immediate threat to anyone, and this fact about his circumstances distinguishes him from those who may legitimately be harmed on the basis the right of self-defense.

But there is an intuitively legitimate theory of crime prevention that is not undermined by hard incompatibilism, nor by other considerations. This view draws an analogy between the treatment of criminals and the treatment of carriers of dangerous diseases. Ferdinand Schoeman (1979) contends that if we have the right to quarantine carriers of severe communicable diseases to protect people, then for the same reason we also have the right to isolate the criminally dangerous. Quarantining someone can be justified when she is not morally responsible for being dangerous to others. If a child is infected with a deadly contagious virus passed on to her prior to birth, quarantine may nevertheless be legitimate. Now suppose that a serial killer continues to pose a grave danger to a community. Even if he is not morally responsible for his crimes, it would be as legitimate to detain him as it is to quarantine a non-responsible carrier of a deadly communicable disease.

It would be morally wrong, however, to treat carriers of the communicable disease more severely than is required to guard against the resulting threat. Similarly, it would be wrong to treat criminals more harshly than is needed to neutralize the danger posed by them. In addition, just as moderately dangerous diseases may only license measures less intrusive than quarantine, so moderately serious criminal tendencies might only justify responses less intrusive than detention. Shoplifting, for example, may warrant only some degree of monitoring. Furthermore, I suspect that a theory modeled on quarantine would not justify measures of the sort whose legitimacy is most in doubt, such as the death penalty or confinement in the worst prisons we have. Moreover, it would demand a degree of concern for the rehabilitation and well-being of the criminal that would decisively alter much of current practice. Just as society has a duty to seek to cure the diseased it quarantines, so it would have a duty to attempt to rehabilitate the criminals it detains. When rehabilitation is impossible, and if protection of society requires indefinite confinement, there would be no justification for making a criminal's life more miserable than the protection of society requires.

## 11 Meaning in Life

Would it be difficult for us to cope without a conception of ourselves as praiseworthy for achieving what makes our lives fulfilled, happy, satisfactory, or worthwhile – for realizing what Ted Honderich calls our *life-hopes*? (Honderich 1988: 382) Honderich contends that there is an aspect of these life-hopes that is undermined by determinism, but that determinism nevertheless leaves them largely intact. I agree, and I develop this position in the following way. First, it is not unreasonable to object that our life-hopes involve an aspiration for praiseworthiness, which the truth of hard

incompatibilism would undermine. Life-hopes are aspirations for achievement, and it might well be that one cannot have an achievement for which one is not also praiseworthy. But then, giving up praiseworthiness would seem to deprive us of our life-hopes altogether. However, achievement and life-hopes are not as closely connected to praiseworthiness as this objection supposes. If an agent hopes for a success in some project, and if she accomplishes what she hoped for, intuitively this outcome would be an achievement of hers even if she is not praiseworthy for it, although at the same time the sense in which it is her achievement may be diminished. For example, if someone hopes that her efforts as a teacher will result in well-educated children, and they do, then there is a clear sense in which she has achieved what she hoped for, even if because she is not in general morally responsible she is not praiseworthy for her efforts.

One might think that hard incompatibilism would instill an attitude of resignation to whatever one's behavioral dispositions together with environmental conditions hold in store. But this isn't clearly right. Even if what we know about our behavioral dispositions and environment gives us reason to believe that our futures will turn out in a particular way, it can often be reasonable to hope that they will turn out differently. For this to be so, it may sometimes be important that we lack complete knowledge of these dispositions and environmental conditions. Imagine that someone reasonably believes that he has a disposition that might well be an impediment to realizing a life-hope. However, because he does not know whether this disposition will in fact have this effect, it remains open for him – that is, it is possible for him for all he knows – that another disposition of his will allow him to transcend the impediment. For instance, imagine that someone aspires to become a successful politician, but is concerned that his fear of public speaking will get in the way. He does not know whether this fear will in fact frustrate his ambition, since it is open for him that he will overcome this problem, perhaps due to a disposition for resolute self-discipline in transcending obstacles of this sort. As a result, he might reasonably hope that he will get over his fear and succeed in his ambition. Given hard incompatibilism, if he in fact does overcome his problem and succeeds in political life, this will not be an achievement of his in as robust a sense as we might naturally suppose, but it will be his achievement in a substantial sense nonetheless.

How significant is the aspect of our life-hopes that we must relinquish on the assumption of hard incompatibilism? Saul Smilansky argues that although determinism leaves room for a limited foundation for the sense of self-worth that derives from achievement or virtue, the hard determinist's (and also, by extension, the hard incompatibilist's) perspective can nevertheless be "extremely damaging to our view of ourselves, to our sense of achievement, worth, and self-respect," especially when it comes to achievement in the

formation of one's own moral character. In response, Smilansky thinks that it would be best for us to foster the illusion that we have free will (Smilansky 1997: 94; 2000). With Smilansky I agree that there is a kind of self-respect that presupposes an incompatibilist foundation, and that it would be threatened if hard determinism or hard incompatibilism were true. I question, however, whether he is right about how damaging it would be for us to give up this sort of self-respect, and whether his appeal to illusion is required.

First, note that our sense of self-worth – our sense that we have value and that are lives are worth living – is to a non-trivial extent due to features not produced by the will, i.e., not voluntarily, at all, let alone by free will. People place great value on natural beauty, native athletic ability, and intelligence, none of which have their source in our volition. To be sure, we also value voluntary efforts – in productive work and altruistic behavior, and indeed, in the formation of moral character. However, does it matter very much to us that these voluntary efforts are also *freely* willed? Perhaps Smilansky overestimates how much we care.

Consider how a person comes to develop a good moral character. It is not implausible that it is formed to a significant degree as a result of upbringing, and, moreover, the belief that this is so is widespread. Parents typically regard themselves as having failed in raising their children if they turn out with immoral dispositions, and parents often take great care to bring their children up to prevent such a result. Accordingly, people often come to believe that they have the good moral character they do largely because they were raised with love and skill. But those who come to believe this about themselves seldom experience dismay because of it. People tend not to become dispirited upon coming to understand that their good moral character is not their own doing, and that they do not deserve a great deal of praise or respect for it. By contrast, they often come to feel more fortunate and thankful. Suppose, however, that there are some who would be overcome with dismay. Would it be justified or even desirable for them to foster the illusion that they nevertheless deserve praise and respect for producing their moral character? I suspect that most would eventually be able to accept the truth without incurring much loss. All of this, I think, would also hold for those who come to believe that they do not deserve praise and respect for producing their moral character because they are not, in general, morally responsible.

## 12 Emotions, Reactive Attitudes, and Personal Relationships

P. F. Strawson (1962) argues that the justification for judgments of blameworthiness and praiseworthiness has its foundation in our reactive attitudes,

emotional reactions to the moral quality of an agent's will. Since moral responsibility, more generally, has this kind of foundation, the truth or falsity of determinism is irrelevant to whether we are justified in regarding agents as morally responsible. These reactive attitudes, such as moral resentment, guilt, gratitude, forgiveness, and love, are required for the kinds of interpersonal relationships that make our lives meaningful. So even if we could give up the reactive attitudes – and Strawson believes that this is impossible – we would never have sufficient practical reason to do so. Thus we would never have sufficient practical reason to cease regarding people as morally responsible. In addition, if determinism did imperil the reactive attitudes, we would face the prospect of the "objective attitude," a cold and calculating stance towards others that would undercut the possibility of meaningful personal relationships.

Strawson is clearly right to believe that an objective attitude would undermine personal relationships, but I deny that we would adopt this stance or that it would be appropriate if we came to accept determinism or hard incompatibilism. In my conception, some of the reactive attitudes would in fact be challenged by hard determinism, or more broadly by hard incompatibilism. I contend that some of these attitudes, such as moral resentment and indignation, presuppose that the person who is the object of the attitude is morally responsible in the "basic desert" sense at issue in the debate. Consequently, such attitudes have presuppositions that the hard incompatibilist would believe to be false. I claim, however, that the reactive attitudes that we would want to retain either are not threatened by hard incompatibilism in this way, or else have analogues or aspects that would not have false presuppositions. The complex of attitudes that would survive does not amount to Strawson's objective attitude, and it would be sufficient to sustain good relationships.

It is plausible that to a certain degree moral resentment and indignation are beyond our power to affect, and thus even supposing that a hard incompatibilist is thoroughly committed to morality and rationality, and that she is admirably in control of her emotional life, she might nevertheless be unable to eradicate these attitudes. Thus as hard incompatibilists we might expect people to be morally resentful in certain circumstances, and we would judge it to be in an important sense beyond the agent's control when they are. However, we also have the ability to prevent, temper, and sometimes to dispel moral resentment, and, given a belief in hard incompatibilism, we might attempt such measures for the sake of morality and rationality. Modifications of this sort, aided by a hard incompatibilist conviction, could well be good for interpersonal relationships.

It might be objected that moral resentment and indignation are crucial to effective communication of wrongdoing in our relationships, and were we to dispel or modify these attitudes, these relationships would be damaged.

However, when someone is wronged in a relationship, she typically has additional attitudes that are not imperiled by hard incompatibilism, whose expression can play the communicative role at issue. These attitudes include being alarmed or distressed about what the other person has done, as well as moral sadness or concern for him. Moral resentment, then, is not clearly required for effective communication in personal relationships.

Forgiveness might seem to presuppose that the person being forgiven is blameworthy, and if this is so, this attitude would also be undermined by hard incompatibilism. But certain key features of forgiveness are not endangered by hard incompatibilism, and they are sufficient to sustain the role forgiveness in its entirety has in relationships. Suppose a friend repeatedly mistreats you, and because of this you resolve to end your relationship with him. However, he then apologizes to you, thereby indicating his recognition that his actions were wrong, his wish that he had not mistreated you, and his genuine commitment to refraining from the offensive behavior. As a result, you decide not to end the friendship. Here the feature of forgiveness that is consistent with hard incompatibilism is the willingness to cease to regard past immoral behavior as a reason to dissolve or weaken a relationship. The aspect of forgiveness that is undercut by hard incompatibilism is the willingness to disregard the friend's blameworthiness. But having given up the belief that we are morally responsible, the hard incompatibilist no longer needs a willingness to disregard blameworthiness in order to enjoy good relationships.

One might argue that hard incompatibilism threatens the self-directed attitudes of guilt and repentance, and that this would be especially bad for relationships. Without guilt and repentance, we would not only be incapable of restoring relationships damaged because we have done wrong, but we would also be kept from retaining our moral integrity. For absent the attitudes of guilt and repentance, we would lack the psychological mechanisms that can play these roles. But note first that it is because guilt essentially involves a belief that one is blameworthy for something that one has done that this attitude would appear to be endangered by hard incompatibilism. It is for this reason that repentance would also seem to be (indirectly) threatened, for feeling guilty would appear to be required to motivate repentance. However, suppose that you behave immorally, but because you endorse hard incompatibilism, you deny that you are blameworthy. Instead, you acknowledge that you have done wrong, you feel sad that you were the agent of wrongdoing, and you deeply regret what you have done (Waller 1990). Also, because you are committed to doing what is right and to your own moral improvement, you resolve to refrain from behavior of this kind in the future, and seek the help of others in sustaining your resolve. None of this is jeopardized by hard incompatibilism.

Gratitude would seem to presuppose that the person to whom one is grateful is morally responsible for a beneficial act, and hard incompatibilism would

then threaten gratitude. However, as in the case of forgiveness, certain aspects of this attitude would be unaffected, and these aspects have the function that gratitude as a whole has in good relationships. Gratitude includes, first of all, being thankful toward a person who has acted beneficially. True, being thankful toward someone typically involves the belief that she is praiseworthy for some action. Still, one can also be thankful to a small child for some kindness, without believing that he is morally responsible. This aspect of thankfulness could be retained even if one gave up the presupposition of praiseworthiness. Often gratitude also involves joy as a response to the beneficent act of another. But no feature of hard incompatibilism undermines being joyful and expressing joy when others are considerate or generous in one's behalf. Expression of joy can bring about the sense of harmony and goodwill often occasioned by gratitude, and thus here hard incompatibilism is not at a disadvantage.

Is the kind of love that mature adults have for each other in good relationships imperiled by hard incompatibilism, as Strawson's line of argument suggests? Consider first whether for loving someone it is important that the person who is loved has and exercises free will in the sense required for moral responsibility. Parents love their children rarely, if ever, for the reason that they possess this sort of free will, or decide to do what is right by free will, or deserve to be loved due to freely-willed choices. Moreover, when adults love each other, it is also very seldom, if at all, for these sorts of reasons. The reasons we love others are surely varied and complex. Besides moral character and behavior, features such as intelligence, appearance, style, and resemblance to certain others in one's personal history all might have a role. Suppose morally admirable qualities are particularly important in occasioning, enriching, and maintaining love. Even if there is an aspect of love that we conceive as a deserved response to morally admirable qualities, it is unlikely that love would be undermined if we came to believe that these qualities are not produced or sustained by freely-willed decisions. Morally admirable qualities are loveable, whether or not people deserve praise for having them.

One might contend that we want to be loved by others as a result of their free will. Against this, the love parents have for their children is typically generated independently of the parents' will altogether, and we don't think that this love is deficient. Kane recognizes this fact about parental love, and he agrees that romantic love is similar in this respect. But he contends that there is a kind of love we very much want that would not exist if all love were determined by factors beyond our control (Kane 1996: 88; cf. Strawson 1986: 309; Anglin 1991: 20). The plausibility of Kane's view might be enhanced by reflecting on how you would react if you discovered that someone you love was causally determined to love you by, say, a benevolent manipulator.

Setting aside *free* will for the moment, when does the will play any role at all in engendering love? When an intimate relationship is disintegrating,

people will sometimes decide to try to restore the love they once had for one another. Or when a student finds herself at odds with a roommate from the outset, she may choose to take steps to make the relationship a good one. When a marriage is arranged, the partners may decide to do what they can to love each other. In such situations we might want others to make a decision that might produce or maintain love. But this is not to say that we would want that decision to be freely willed in the sense required for moral responsibility. For it is not clear that value would be added by the decision's being free in this sense. Moreover, although in some circumstances we might want others to make decisions of this sort, we would typically prefer love that did not require such choices. This is so not only for intimate romantic relationships – where it is quite obvious – but also for friendships and relationships between parents and children.

Suppose Kane's view could be defended, and we did want love that is freely willed in the sense required for moral responsibility. If we in fact desired love of this kind, then we would want a kind of love that would be impossible if hard incompatibilism were true. Still, the sorts of love not challenged by hard incompatibilism are sufficient for good relationships. If we can aspire to the kind of love parents typically have for their children, or the type romantic lovers share, or the sort had by friends who were immediately attracted to each other, and whose friendship became close through their interactions, then the possibility of fulfillment through interpersonal relationships remains intact.

Hard incompatibilism, therefore, does not yield a threat to interpersonal relationships, although it might challenge certain attitudes that typically have a role in such relationships. Moral resentment, indignation, and guilt would likely be irrational for a hard incompatibilist, since these attitudes would have presuppositions believed to be false. But these attitudes are either not required for good relationships, or they have analogues that could play their typical role. Moreover, love – the attitude most essential to good interpersonal relationships – does not seem threatened by hard incompatibilism at all. Love of another involves, fundamentally, wishing for the other's good, taking on her aims and desires, and a desire to be together with her, and none of this is endangered by hard incompatibilism.

## 13 Revisionism

Manuel Vargas (2005) points out that theories of free will and moral responsibility can be revisionist about our ordinary conceptions, attitudes, and practices to various degrees. The version of hard incompatibilism I endorse is strongly revisionist about the key notion of moral responsibility. Our ordinary conception has it that people can be morally responsible in the "basic

desert" sense, but this conception is in error. At the same time, I think that people are morally responsible in other senses. But giving up the "basic desert" notion of responsibility while retaining these other senses would amount to a considerable revision. Its magnitude would be evident in the resulting radical change in how we would assess moral resentment and indignation. We ordinarily think that these attitudes are often fully justified, while on the hard incompatibilist view, they always presuppose the false belief that those toward whom they are directed are blameworthy. If these attitudes were merely sporadic, then hard incompatibilism's revisionist view of these attitudes might still count as weak or merely moderate. But they are not sporadic; for example, most people are morally resentful of someone or other most of the time. With regard to treatment of criminals, the degree to which my view is revisionist is relative to specific practices and justifications for them. About the reactive attitudes required for good interpersonal relationships my position is merely weakly or moderately revisionist. Moral sadness, gratitude, forgiveness, and love are not endangered by my version of hard incompatibilism, or else have analogues or aspects that would not have false presuppositions. Consequently, for these attitudes at most only a fairly mild sort of revision would be required.

## 14 The Good in Hard Incompatibilism

Hard incompatibilism also promises substantial benefits for human life. Of all the attitudes associated with the assumption that we are morally responsible, anger seems most closely connected with it. Discussions about moral responsibility most often focus not on how we regard morally exemplary agents, but rather on how we consider those who are morally deficient. Examples designed to elicit a strong intuition that an agent is morally responsible typically feature an especially malevolent action, and the intuition usually involves sympathetic anger. It may be, then, that our attachment to the assumption that we are morally responsible derives to a significant degree from the role anger plays in our emotional lives. Perhaps we are disposed to feel that giving up this assumption of responsibility poses a serious threat because the rationality of anger would be undermined as a result.

The type of anger at issue is the kind that is directed toward someone who is believed to have behaved immorally – it comprises both moral resentment and indignation. Let us call this attitude *moral anger*. Not all anger is moral anger. One type of non-moral anger is directed at someone because his abilities are lacking in some respect or because he has performed poorly in some situation. At times our anger is without object. Still, most human anger is moral anger.

Moral anger comprises an important part of our moral lives as we ordinarily conceive them. It motivates us to resist abuse, discrimination, and oppression. At the same time, expression of moral anger frequently has harmful effects, failing to contribute to the well-being either of those to whom it is directed or of those expressing the anger. Often its expression is intended to cause little else than emotional or physical pain. As a result, it has a tendency to damage relationships, impair the functioning of organizations, and unsettle societies. In extreme cases, it can motivate people to torture and kill.

The realization that expression of moral anger can be damaging gives rise to a strong demand that it be morally justified when it is to occur. The demand to morally justify behavior that is harmful is generally a strong one, and expressions of moral anger are typically harmful. This demand is made more urgent by the fact that we are often attached to moral anger, and that we not infrequently enjoy expressing it. Most commonly we justify expression of moral anger by contending that wrongdoers deserve it, and we presume that they deserve it in the basic sense.

On the assumption of hard incompatibilism, however, justification of this sort is not available. But given the concerns to which expression of moral anger give rise, this may be a good thing.

Accepting hard incompatibilism is unlikely to modify our dispositions to the extent that expression of moral anger ceases to be a problem for us. At the same time, note that moral anger is often sustained and magnified by the belief that its object is morally responsible for immoral behavior. Destructive moral anger in relationships is fostered in this way by the assumption that the other is blameworthy. The anger that fuels ethnic conflicts is nourished by the supposition that a group of people deserves blame for past wrongs. Hard incompatibilism advocates giving up such beliefs because they are false. As a consequence, moral anger might decrease, and its expressions subside.

Would the benefits that would result if moral anger were curtailed in this way compensate for the losses that would ensue? Moral anger motivates us to oppose immoral behavior. Would this benefit be relinquished? If for hard incompatibilist reasons the assumption that wrongdoers are blameworthy is withdrawn, the conviction that they have in fact behaved immorally would not be threatened. Even if those who perpetrate genocide are not morally responsible, their actions are nonetheless clearly immoral, and a belief that this is so would remain untouched. This, together with a commitment to oppose wrongdoing, would allow for a resolve to resist abuse, discrimination, and oppression. Accepting hard incompatibilism would thus permit us to retain the benefits moral anger can also provide, while at the same time challenging its destructive effects.

## Further Reading

The ideas in this chapter are presented and discussed in greater detail in my *Living Without Free Will* (Cambridge: Cambridge University Press, 2001). This chapter features some key improvements over the book, especially in the argument for source and against leeway views, and in the treatment of event-causal libertarianism.

As mentioned in the chapter, Baruch Spinoza argues for hard determinism in his *Ethics*, first published just after his death in 1677; the current standard translation is in *The Collected Works of Spinoza*, ed. and tr. Edwin Curley, volume 1 (Princeton, NJ: Princeton University Press, 1985). About a century later hard determinism is defended by Baron Paul d'Holbach in his *Système de la Nature* (Amsterdam, 1770), and by Joseph Priestley (who also made important contributions to modern chemistry) in *A Free Discussion of the Doctrines of Materialism and Philosophical Necessity, In a Correspondence between Dr. Price and Dr. Priestley* (1788), reprinted in Joseph Priestley, *Priestley's Writings on Philosophy, Science, and Politics*, ed. John Passmore (New York: Collier, 1965).

The view that morally responsibility is in fact impossible – whether determinism or indeterminism is true – is defended by Galen Strawson in *Freedom and Belief* (Oxford: Oxford University Press, 1986), and in a number of more recent articles, including "The Impossibility of Moral Responsibility," *Philosophical Studies* 75 (1994), 5–24. Saul Smilansky, in *Free Will and Illusion* (Oxford: Oxford University Press, 2000), endorses an argument of the sort that Strawson advocates, but goes on to contend that for us to believe that we lack the sort of free will required for moral responsibility would be harmful, and thus it would be best to maintain the illusion that we have this kind of free will. Detailed versions of hard determinism or hard incompatibilism have been presented by Ted Honderich in *A Theory of Determinism* (Oxford: Oxford University Press, 1988), and by Bruce Waller in *Freedom Without Responsibility* (Philadelphia: Temple University Press, 1990).

Richard Double, in *The Non-Reality of Free Will* (New York: Oxford University Press, 1991), and in *Metaphilosophy and Free Will* (New York: Oxford University Press, 1996), argues that the claim that we have the free will required for moral responsibility cannot be true for the reason that the very concept of free will is internally incoherent if it is construed in a realist, non-subjectivist way.

# 4
# *Revisionism*

*Manuel Vargas*

> [I]t goes without saying that I do not deny, presupposing I am no fool, that many actions called immoral ought to be avoided and resisted, or that many actions called moral ought to be done and encouraged – but *for different reasons than formerly.*
>
> *F. Nietzsche,* Daybreak *§103*

> The ethical commonplaces of any period include ideas that may have been radical discoveries in a previous age. This is true of modern conceptions of liberty, equality, and democracy, and we are in the midst of ethical debates that will probably result two hundred years hence in a disseminated moral sensibility that people of our time would likely find very unfamiliar.
>
> *T. Nagel, "Ethics Without Biology," in* Mortal Questions, *p. 143*

## 1 A Brief History of Some Concepts

Consider the history of thinking about three different things: water, marriage, and being a magician.

At one point in time people thought water was one of the four basic indivisible substances of the universe. After the hard work of a number of people, we eventually figured out that water is $H_2O$, something that was neither indivisible nor as basic as some had thought. When we learned that water was not one the four basic indivisible substances of the world, we did not collectively curse the discovery that water did not exist. Instead, we proceeded as before with the additional bit of knowledge that water was somewhat different than we had thought.

Now think about marriage. Throughout much of human history marriage was thought of as a property transaction. In some places I would give you eight camels and some wine in exchange for your daughter. In other regions of the world you would give me an ox and some money if I would take your

sister off your hands. When it was all done, I had a new piece of property that I was free to treat as I wished. We called that new piece of property a "wife" and if I was especially wealthy or had come to inherit other similar pieces of property, I could even have multiple wives. In many places customs eventually changed. We stopped thinking of wives as pieces of property and we soon stopped thinking of marriages as property transactions. We did not stop getting married, however. Instead, we simply decided that marriage did not have to rest on the assumptions of the bad old days: marriage could change in light of our growing recognition that women were – *shockingly enough* – not property.

One last example: in many cultures and places people called someone a magician (or something similar in the local language) if they believed that person could cast spells or otherwise had magical abilities. However, today when David Blaine or David Copperfield announces a performance of a magic show, we do not feel robbed that we did not witness a demonstration of occult powers. We do not threaten the Society of American Magicians or the International Brotherhood of Magicians with a lawsuit for false advertising. Instead, we understand that when people talk about Blaine, Copperfield, and so on as magicians, we understand that they mean people who create illusions that have the appearance of violating laws of nature. And, I suspect, few if any adults in their audiences suppose that their magic consists in the invocation of supernatural forces.

In all three of these cases we came to change how we thought about the nature of these things (water, marriage, and magicians), without thereby concluding that water did not exist, that no one had ever been married, and that there were no magicians. These changes did not happen by themselves. They were all driven by difficulties concerning older conceptions of these things. As we learned more about the world and about ourselves, it made sense to acknowledge that how we had previously thought about these things was mistaken. Crucially, the mistakes weren't fatal. In each of these cases we continued to use the revised concept, but in a different and better way. Revisionism about free will and moral responsibility is the idea that we should do something similar for how we think about free will and moral responsibility. In a nutshell, revisionism is the view that what we ought to believe about free will and moral responsibility is different than what we tend to think about these things.

In this chapter I will say more about why revisionism about free will is called for, what it amounts to, how it is different than other views, what version of it I favor, and why it is the most promising solution to that cluster of problems philosophers argue about under the heading of "free will." To foreshadow what I will argue for: We tend to think of ourselves as having a powerful kind of agency, of the sort described by various libertarian accounts.

That is, we see ourselves as having genuine, robust alternative possibilities available to us at various moments of decision. We may even see ourselves as agent-causes, a special kind of cause distinct from the non-agential parts of the causal order. Moreover, we tend to think of this picture of our own agency as underwriting many important aspects of human life, including moral responsibility. How we think about a range of social issues (crime and punishment, addiction, and even issues such as homelessness), and the social policies we construct around them, in part depend on the presumption of this picture of agency. The problem is that our self-conception is implausible and largely unnecessary. It requires a metaphysics of agency that we have no independent reason to believe in and it mistakenly holds that we cannot attain a range of important human and moral aspects of our life in its absence. What I will argue is that we can get by with a stripped-down conception of agency that avoids many of the problems that plague our libertarian self-conception. It does, however, require some revision in how we think about ourselves and how we understand the foundations of various moral practices.

Importantly, revisionism does not, by itself, require that we jettison talk of moral responsibility, praise, and blame. Revisionism does not commit us to dismissing the pull of incompatibilist construals of our self-conception. Revisionism does not require that we deny that there is something right about compatibilist and libertarian claims that we are free and responsible. What revisionism does require is that we regard our intuitive, commonsense self-conception with a critical eye, giving up those parts that are least plausible or otherwise worth abandoning.

## 2 Building a Theory of Free Will

It is not clear that there is any single thing that people have had in mind by the term "free will." Perhaps the dominant characterization in the history of philosophy is that it is something like the freedom condition on moral responsibility. Roughly, the idea is that to be morally responsible for something, you had to have some amount of freedom, at some suitable time prior to the action or outcome for which you are responsible. That sense of freedom – whatever it amounts to – is what we mean to get at by the phrase "free will." However, there may be things for which free will might be important or other senses of free will that are independent of concerns about moral responsibility. For example, philosophers have worried whether free will is required for some human achievements to have a special worth or value, or for there to be values and valuing in any robust sense. Although I think much of what I will say can be applied to other aspects of thinking about it, I will primarily be

concerned with free will in its connection to moral responsibility, the sense in which people are appropriately praised or blamed.

Although it might seem that the first task for a philosophical account of free will is to begin with a characterization of it, there is a prior issue that needs clarification. The prior issue concerns the sort of account of free will that we intend to offer. We could offer an account of free will that attempts to reflect how we tend to think about and talk about free will. The success of this sort of account turns on how faithfully it captures and fills in the details of the way we tend to think and talk about free will. Call this sort of account a *diagnostic* account – it attempts to describe how we do, in fact, think about free will. It might fill in additional details beyond how people commonly think about the issue, but these additional details would be constrained by the contours of commonsense. These philosophical details would be like elaborate carvings on a piece of wood whose basic shape must be preserved. In contrast, we could offer an account that is not principally concerned to respond to our ordinary intuitions or commonsense thinking about free will, and that instead aims to tell us how we ought think about it. Call this other sort of account *prescriptive*.

An ideal account might be able to offer a comprehensive theory of free will that is both diagnostic and prescriptive. (I say diagnostic rather than descriptive to avoid confusion about mixed cases: to call someone a "cheap-ass bastard" might be both descriptive and prescriptive. Since not much hangs on this, feel free to read "diagnostic" as "descriptive" if you like.) A great deal of the philosophical literature on free will assumes that a comprehensive account must be a unified account – that is, that both the diagnosis and the prescription are the same. For example, someone might offer a diagnosis of commonsense that argues for some compatibilist characterization of free will (for example, a person acts freely when they act in a way consistent with what they value). The same philosopher might then go on to offer the prescription (usually implicit) that this is how we ought to conceive of free will.

It is not surprising that we assume that our default conception of things is correct. The trouble is there have been plenty of cases where we learned that this was simply not true (for example, water). In other cases we came to have reason to change how we thought about something (marriage, magicians) through a somewhat different route. For now, the point to bear in mind is simply that we should not suppose that a comprehensive account of free will must offer the same prescription as it offers in the analysis of our concept. To reflect this possibility, my account will come in two stages: diagnosis and prescription. In the diagnostic part I will argue that our ordinary thinking about free will has elements that are incompatibilist. In the prescriptive part I will argue that we should revise away from these commitments.

*An aside on philosophical methodology*

Sometimes philosophers offer accounts of free will that are not intended to cohere neatly with commonsense but are instead intended as "paraphrasings" or "cleaned up" accounts of our commonsense notions. Oftentimes the motivation for this approach is the recognition that commonsense thinking can be very messy. The aim of a paraphrasing approach is to abstract away from the occasionally inconsistent details of ordinary thinking and to provide a coherent, systematic account that constitutes something of a repair to common sense. Such an account diverges in various ways from our pre-philosophical conceptions of things, but these divergences are typically treated as minor or unremarkable.

Although I share some of the spirit of this methodological approach, I am inclined to think that its usual method of implementation brings with it a serious cost. In particular, paraphrasing accounts run the risk of sliding back and forth between raw description and a proposal for philosophical repair of commonsense (a similar concern is expressed in Nichols and Knobe, forthcoming). Part of the risk is that this sliding between description and repair might be *ad hoc* or insufficiently principled or motivated. Even when it is principled, though, a further difficulty arises when we try to keep straight the argumentative burdens faced by the proposal. For example, when something is proposed as a repair to commonsense, we needn't worry about intuitions that rely on aspects of the description that the repair has abandoned. However, intuitive plausibility *is* important when what is at stake is a description or diagnosis of commonsense. Paraphrasing accounts can make it unclear when something is intended as a repair and when it is not. In turn, this invites philosophers to talk past each other, with incompatibilist ships of descriptions passing compatibilist ships of repair in the philosophical night.

The above characterization is not without its shortcomings (e.g., a good many compatibilists would resist the construal of their views as repairs to commonsense thinking). It does, however, capture one thing that can happen when methodological assumptions are left unstated.

To the extent that there are distinctive methodological approaches in the free will debate, I suspect that one source is rooted in the distinctive methodologies that dominate fields that adjoin the free will problem: metaphysics and ethics. In the latter, tolerance of some degree of revision away from commonsense beliefs has been the order of the day for a long time. It received a particularly influential statement in the work of Rawls and the method of reflective equilibrium (Rawls, 1971). In contrast, contemporary metaphysics has, at least methodologically, hewn more closely to approaches that less readily forsake the constraints of commonsense, something like what P.F. Strawson called "descriptive metaphysics" or "concept-mapping" (Strawson,

1992). This characterization oversimplifies the methodological diversity in both metaphysics and ethics, but it does begin to explain why differences between those who think the free will problem is primarily metaphysical and those who think of it as primarily normative might lead each group to think that the other has missed the point in their respective treatments of the problem.

## 3 The Case for an Incompatibilist Diagnosis of Commonsense

Here I begin the case for thinking that commonsense is incompatibilist, that is, that commonsense has elements that require a picture of agency whose commitments could not be satisfied in a deterministic world.

I will focus on three broad lines of argument, each of which favors the conclusion that commonsense is incompatibilist. First, there are considerations grounded on the traditional philosophical arguments for incompatibilism. Second, there is a range of experimental data that suggests that ordinary thinking about free will and moral responsibility are at least partly incompatibilist. Third, reflections on cultural and social history also seem to favor incompatibilism.

None of these considerations are decisive when taken individually. I doubt there are any decisive considerations to be had here or in many areas of philosophy. That said, I do think that the considerations I suggest, especially when taken as a group, strongly favor an incompatibilist diagnosis of commonsense thinking about free will and moral responsibility.

### *Traditional philosophical arguments for incompatibilism*

Surely the most important argument for incompatibilism over the past few decades has been a family of arguments that include Peter van Inwagen's much-discussed Consequence Argument (van Inwagen, 1983). The other authors have discussed this argument in some detail, so I'll be brief about what I take to be its salience. The core idea of this way of arguing for incompatibilism begins by reflecting on the idea that if determinism is true, there is only one way, physically speaking, the world can turn out if we hold fixed the past and the laws of nature. Thus, for someone to do otherwise in a deterministic world, the agent would have to do something that results either in a change of the laws of physics or requires some difference in the past. Since changing the laws and changing the past do not look possible for anyone, then in a deterministic world no one would have the ability to do

otherwise. Thus, if determinism is true then we lack the ability to do otherwise. When this is connected with a principle that free will (and/or moral responsibility) requires the ability to do otherwise, we get the basic shape of perhaps the most influential approach to incompatibilism.

There are a number of ways to resist this path to incompatibilism. One is to challenge the idea that free will does not require the ability to do otherwise. Another is to contend that that the ability to do otherwise required by free will is not the sense of the ability to do otherwise that is ruled out by the argument. Another is to dispute one or another of the principles of reasoning in the argument. Yet another way is to grant the conclusion, but to argue that, whatever is the case for free will, *moral responsibility* does not require alternative possibilities of the sort that are incompatible with determinism (see John Martin Fischer's view in chapter 2). There is an impressively large and sophisticated literature on all of these possibilities, and canvassing them all would be impossible to do in this chapter. Instead, I will largely focus on developing an underappreciated aspect of these arguments.

I am somewhat more optimistic than Robert Kane is about the results of these arguments favoring incompatibilists (see chapter 1). I think the dialectical situation is a bit like fighting the hydra of mythology: for every difficulty found with these arguments, two new arguments emerge. (Well perhaps not *two*, but certainly another.) What I will focus on, though, is not the contents of the argument directly, but rather a different kind of argument, one that relies on the contents of traditional arguments only indirectly. To see what I have in mind, it helps to think about *why* arguments such as the Consequence Argument for incompatibilism have been influential. What makes these arguments powerful is not so much that they rule out the possibility of compatibilism but rather that they show how easily incompatibilism seems to capture ordinary ways of thinking about our own agency.

To see why this relatively innocuous point matters, consider perhaps the most influential line of criticism against the Consequence Argument and its successors (henceforth, Consequence-style arguments). Several critics have focused on how an antecedently compatibilist reading of the relevant ability term (for example, the ability to break the laws of nature) makes the argument unpersuasive against antecedently committed compatibilists. Suppose these critics are right. If so, it would mean that Consequence-style arguments couldn't rule out the possibility of compatibilism. From this, we might think things look like a standoff, at best, and at worst a real problem for incompatibilism. Here, the innocuous issue has some role to play: the "naturalness" or ease of the incompatibilist reading of the argument is itself evidence that the argument captures an important part of the contents and logic of commonsense thinking about these issues. Even if the contents of the argument cannot rule out a compatibilist reading, the naturalness of the incompatibilist

readings of the argument *strongly* suggests that we can and do understand these issues in incompatibilist ways, at least sometimes. Try it – present the Consequence Argument or its successor to a group of people and see how many people read it the first time (or even the fifth time!) in a way that exploits a compatibilist interpretation of "can." You won't find many people, which in turn suggests that the default way of thinking about these issues really is incompatibilist.

If a compatibilist construal of our theoretical commitments on the freedom-relevant notion of ability were front and center in our deliberations about free will and moral responsibility, it is difficult to see how Consequence-style arguments for incompatibilism would have been as persuasive as they have been to so many people. More to the point, if we didn't have incompatibilist intuitions, it is hard to see how the argument could have ever seemed compelling. So, even if the argument doesn't directly prove that commonsense is incompatibilist (for we might discover that any such argument can be read in a compatibilist way), it indirectly provides evidence for incompatibilism by being so widely and easily read in incompatibilist terms.

Although this approach to incompatibilism has its limitations I think that the Consequence Argument and its progeny have done a good job of making manifest an important and natural understanding of the sense of ability relevant to questions of free will. It seems to me that it puts the burden on compatibilists to show that a similar or better mustering of our intuitions and concepts can be done on behalf of compatibilism.

On this construal of the significance of incompatibilist arguments, the issue becomes whether alternative (and compatibilist) construals of ability are the ones we have in mind when discussing these issues. The absence of anything like agreement by compatibilist (and other) critics on how to construe the notion of ability or "can" in Consequence-style arguments suggests that even if there are alternative ways to read the argument, there is no systematic logic or thread to our thinking that is *more* uniform and natural than the incompatibilist construal of the argument.

One way to respond to my construal of the power of Consequence-style arguments is to hold that compatibilists have better, or at least equally good tools to show that our ordinary thinking is fundamentally compatibilist in its commitments. Recall the path to incompatibilism we have been considering works through an argument that determinism rules out the ability to do otherwise plus a principle that the ability to do otherwise is required for moral responsibility. A family of thought experiments that have become known as "Frankfurt cases" could be taken to show a real shortcoming of this approach to incompatibilism by showing the failure of this last step, the step that holds that alternative possibilities are required for moral responsibility.

The basic idea in a Frankfurt-style counterexample to a principle that holds that moral responsibility requires alternative possibilities is illustrated in this scenario:

> There is an agent facing a choice, where unbeknownst to that agent an intervener is prepared to induce some condition that brings it about that the agent make a particular choice (call it the Bad Choice), should the agent fail to make the choice on his or her own. Nevertheless, the agent makes the Bad Choice on his or her own, and the intervener never acts.

What this scenario is supposed to illustrate is that an agent can appropriately be held responsible for some choice or outcome, even if he or she lacked alternative possibilities. If true, this would show that the last piece of the traditional incompatibilist argument we have been considering would be mistaken. Frankfurt-style cases (from here on, simply "Frankfurt cases") block the last step of the argument, and in turn, they seem to suggest that our intuitions about at least the responsibility-relevant sense of ability do not require alternative possibilities.

Again, the complexity of these issues outstrips the constraints of this chapter, but there are a few simple lessons we can draw about these issues. (See, however the prior chapters for some dissenting views about what lessons should be drawn.)

First, it is clear that first generation Frankfurt-cases don't work. Frankfurt's original cases seem to rely on having under-described the case. If we stipulate from the start that the scenario is deterministic it isn't obvious that we should view it as one in which there is moral responsibility. If the scenario is indeterministic, then it looks like there are alternative possibilities in the case: it is merely their full realization in action that it is prevented by the presence of the intervener. (That is, if it is indeterministic, it looks like the agent has the ability to make a different choice, or at least to start to make a different choice – for example, the Good Choice – forcing the intervener to intervene.) Call this *the dilemma strategy*.

Something like the dilemma strategy defuses a good number of Frankfurt cases, beyond Frankfurt's original example. There are cases where the dilemma strategy seems less compelling, though. In some examples, the alternatives can seem so feeble or disconnected from the context of choice that it seems strange to think that this is the sort of thing required for moral responsibility. Suppose the only alternative to the choice is something really bizarre (use your imagination). Would the bare presence of this alternative – irrespective of how strange it might be – be sufficient for responsibility? Suppose you are trying to decide whom to vote for and there is only one possibility open to you. On the dilemma strategy, we would ordinarily hold that you are not

responsible. But suppose that you suddenly gain an additional possibility – to, say, roast a chicken, as Michael McKenna (2003) has suggested. Why would the addition of this possibility suddenly turn you into a responsible agent? If lack of alternative possibilities is what makes someone not responsible in a Frankfurt-case, why would the addition of strange or trivial possibility suddenly make an agent responsible? How could roasting a chicken (or whatever bizarre alternative you imagined) be at all relevant to whether or not there is moral responsibility in the scenario? Call this response to the dilemma strategy *the normative relevance objection*.

While there are a number of things that critics of Frankfurt-cases might say at this point, I want to note something strange about the introduction of the normative relevance objection. It is not clear why it is relevant. What we have set out to do is to characterize the web of folk concepts concerning free will and moral responsibility. As I suggested at the start of this chapter, it would be a mistake to suppose that our thinking, even about free will and moral responsibility, is flawless and well supported. Indeed, if the history of moral thinking is any indication, it is especially true of our widely held moral values that we have a long history of coming to decide that they are or have been mistaken. What the normative relevance objection presupposes is that any demand for alternative possibilities should be normatively relevant to our deciding that someone is morally responsible. I see no reason why we should accept this supposition. One thing we might discover is that we really do have an alternative possibilities requirement in our thinking about responsibility, but that it does not do any justifiable work. (So, perhaps recent Frankfurt-style cases show this much.) It is therefore open to the incompatibilist to hold that we do have an alternative possibilities requirement in our ordinary thinking about responsibility while admitting that this requirement may, at least sometimes, be normatively spurious. This admission might invite concerns about whether there is reason to preserve a sometimes-spurious requirement, but those are concerns for a later stage in our theorizing and not appropriate for the project of a diagnostic account of the folk concepts of free will and moral responsibility.

A different sort of strategy is available to the proponent of Frankfurt-style cases: one could endeavor to build more elaborate versions of the case that duck the various objections that have been lodged against earlier versions of the case. The growing number of epicycles on these cases (e.g., involving parallel brain processes, one indeterministic, the other deterministic) can do important work answering objections to less complex Frankfurt-cases. The rapidly increasing complexity of these cases also comes at a cost: it makes it harder to get a clear-headed assessment of these cases as evidence that we do not ordinarily require alternative possibilities in our assessments of responsibility. Inasmuch as we are plumbing the depths of our folk concepts of

freedom and responsibility sufficiently complex cases run the risk of testing only the intuitions of philosophers with well-cultivated and deeply entrenched views about the matter. Moreover, the more a particular case relies upon mechanisms remote from our ordinary and perhaps naïve understanding of human agency the more likely it is that our commonsense understandings of these things will get distorted by the mechanisms imagined in the case. Although these issues are not resolved, I remain skeptical (perhaps more so than most) that we will be able to show that alternative possibilities aren't a deep and pervasive part of thinking about freedom and responsibility.

As both Kane and Pereboom have noted, there is another route to incompatibilism that does not obviously rely on alternative possibilities and the ability to do otherwise. Instead, it focuses on the idea that an action has to be ultimately up to the agent in some sense. What these incompatibilists – *source incompatibilists* – go on to argue is that this sense of ultimacy is not the sort of thing that is compatible with determinism. This route to incompatibilism has the advantage of sidestepping Frankfurt cases and the complex web of issues they raise. I am uncertain about how pervasive incompatibilist intuitions of ultimacy (and thus, source incompatibilism) are among non-philosophers, although I am open to the possibility that the commonsense concept of moral responsibility does have this element. I suspect that it is there in different degrees in different people, but somewhat less widespread than alternative possibilities intuitions. Even so, it may be widespread enough to merit counting as part of the folk conceptions of free will and moral responsibility.

## *Experimental evidence in favor of an incompatibilist diagnosis of folk concepts*

I believe that the traditional philosophical arguments generally favor an incompatibilist and alternative possibilities reading of our commonsense requirements for free will and moral responsibility. However, I think that there is a potentially more powerful way to show that commonsense thinking – what psychologists and others sometimes call "folk" thinking – is incompatibilist. We can examine experimental data. (For what it is worth, the label "folk" isn't supposed to be derogatory – it just refers to what we might think of as "ordinary folks.") I see no reason to think that philosophers are uniquely or even especially well equipped to determine the contents of commonsense beliefs. Indeed, the power of philosophical arguments about free will seems to rest on their intuitive force, on their ability to capture folk thinking about abilities, the meaning of *can*, and so on. If experimental data can tell us something about these things, then we should pay attention.

As it turns out, a number of psychologists and empirically oriented philosophers have been doing experimental work relevant to these issues. One especially interesting set of results come from the work of Shaun Nichols and Joshua Knobe. In one experiment, they gave their subjects descriptions of two different universes, one in which everything is completely caused by whatever happened before it, and the other a universe in which almost everything is determined by whatever happened before it, *except* human decision making. Then, they asked their subjects to identify which universe is more like ours. Roughly 95 percent of respondents describe the second universe (the one in which human decision making was indeterministic) as the one most like ours. This result seems to strongly favor the view that our ordinary self-conception of human agency is incompatibilist (specifically, libertarian). It is difficult to imagine why we would suppose human decision making is exempt from determinism if it were not linked to our having free will.

Interestingly, however, compatibilist intuitions can be elicited in different contexts. A number of experiments, show that subjects tend to give compatibilist responses (that is, they ascribe freedom and responsibility even when they are told the world is deterministic) in cases where the example is concrete and triggers strong emotional reactions (Nahmias et al., forthcoming; Nichols, 2006; Nichols and Knobe, 2006; Woolfolk et al., 2006). Incompatibilist reactions never disappear entirely – they survive at rates roughly between a quarter to a third of respondents. Nevertheless, we also seem to sometimes act as though our ordinary concept were compatibilist.

What is going on? It would take us far a field to pursue the possibility that one or more studies are methodologically flawed or that the data are anomalous. I am inclined to think the data are good enough that we are better off assuming the methodological soundness and the consistency of the results. In light of this there are two things we might conclude.

First, we might conclude that our ordinary concepts of free will and moral responsibility are not unified. Perhaps we inconsistently deploy different concepts of freedom and responsibility. If so, this would still be something of a victory for incompatibilists. Incompatibilists need not – indeed, generally do not – deny that there are conditions under which we utilize compatibilist notions of freedom. What incompatibilists about the folk concept must hold is that our ascriptions of free will and moral responsibility are incompatible with determinism being true in some important sense. For compatibilism to be a meaningful position in this context, it must hold that there is *no* important sense in which our ascriptions of free will and moral responsibility are incompatible with the thesis of determinism. What the experimental data appear to show is that we really do imagine ourselves to be agents

with genuine, metaphysically robust alternative possibilities, and we really do, at least in moments of cool, abstract consideration, tend to favor an alternative possibilities requirement on moral responsibility. So, the experimental data seem to be something of a victory for incompatibilist diagnoses of commonsense.

Second, we could read the empirical data in a different and perhaps complementary way. We might hold that what the experiments are illustrating is the difference between, on the one hand, our genuine theoretical beliefs about free will and moral responsibility, and, on the other, the pragmatic dimension of holding people responsible. When the assignment of praise, blame, and punishment to a particular individual is not at stake, our reactions manifest a conception of things that seems straightforwardly incompatibilist. However, when the assignment of praise and blame for a particular individual (fictional or real) becomes a live possibility, this triggers pragmatic considerations about the importance of holding people responsible that may swamp whatever more nuanced requirements we have for free will and moral responsibility. In general, there are good pragmatic reasons to assume that people are responsible agents unless we have reason to think otherwise, and we need a pretty good reason to *not* assign responsibility to people unfamiliar to us. Systems of sanctioning (punishment, both formal and informal) are simply too important. Compatibilist judgments reflect the pragmatic dimensions of a socially embedded practice whose efficacy turns, in part, on swift responses to harm.

Either way, these results look like evidence that our ordinary understanding of free will and moral responsibility have incompatibilist commitments, and that any theory that fails to acknowledge this will fail as a diagnostic theory of our folk concepts.

The experimental results I have alluded to may yet be overturned. Moreover, the issue of exactly what chunk of society has or is committed to a libertarian self-conception and how culturally widespread these convictions are remains unsettled. The currently available evidence clearly suggests that a good percentage of us have incompatibilist commitments in our conception of our agency. This is compatible with it turning out that there are others of us altogether lacking in incompatibilist inclination. Nevertheless, even if it turns out (despite current evidence) that the majority of people in, say, the United States do *not* have a conception of agency that is libertarian, it might still remain plausible to think that many of our social institutions and social policies were generated in a context where people *did* have those commitments, and where those commitments informed how those social institutions and practices developed. This is something I will turn to consider; powerful ideas take on a life of their own, and they can structure our thinking long after we regard them as unwarranted or mistaken.

### *Cultural history and incompatibilism*

Before turning to the prescriptive account, it may be worth asking how we came to have a libertarian self-image, to the extent that we do. There is no single widely accepted answer to this question. Undoubtedly, there are multiple sources of our self-image, sometimes in tension with one another and sometimes mutually supporting. One possibility suggested by several philosophers is that we have innate, evolved psychological mechanisms whose role in responsibility ascriptions gives rise to our libertarian self-conception. Another possible source is the tendency to infer libertarianism from our first-person phenomenological experience. Irrespective of whether or not we are *warranted* in drawing any inferences about our freedom from our phenomenology of deliberation and decision-making, it can sure seem like we have genuinely open alternative possibilities available to us when we are deciding what to do. This experience, ubiquitous as it is, might fuel the belief that those seemings are real. A third possible source of the incompatibilist intuitions we tend to have may be rooted in the cultural history of the West.

There is a long tradition of dualism in the western world, stretching back to at least Plato. If you accept that the mind or soul is fundamentally different than the physical world, it will be natural to have a conception of agency that is not governed by physical laws. Given the phenomenology of decision-making, and given a conception of agency ungoverned by physical laws, a libertarian conception of agency might seem reasonable. Moreover, the legacy of mainline Christianity may be relevant. Acceptance of dualism in the history of Christianity is significant. Relatedly, libertarianism was (and continues to be) important to many of the most influential figures in this history of Christian thought. For theologians and philosophers, libertarianism often appears to be the best hope to explain how a loving, omniscient, and omnipotent God could permit so many harms in the world. If there is some great value in libertarian freedom, and if we do have libertarian freedom, it allows us to make sense of at least some harms that God might have otherwise prevented. There are, of course, threads of the Christian intellectual tradition that cut against some of these considerations. Still, between dualism, the problem of evil, and what centuries may have translated into the concerns of Sunday sermons, there has been, plausibly, a web of cultural reinforcement for a libertarian picture of our agency.

What should we think about the relevance of cultural history? We *could* suppose that our default, commonsense view of free will developed independent of these forces and is in no way part of the cultural heritage bequeathed to us by the contingent history of the West. But this strikes me as naïve and wholly unrealistic. At the very least, we should be open to the idea that some of our ordinary commitment to incompatibilism is at least reinforced by – if

not rooted in – aspects of our contingent cultural history. Irrespective of what one thinks about the truth of dualism and the influence of religious beliefs, professional philosophers (who are overwhelmingly non-religious) sometimes need to be reminded of the powerful grip these pictures have on many people, especially in the United States. Even if no one were religious today, the fact of the West's religious history would be relevant. As Nietzsche once noted, the power of a religion's metaphysical and moral picture can persist long after religiosity is perceived as embarrassing, antiquated, or irrelevant to secular life (Nietzsche, 1996 [1887], §9).

This concludes my discussion of the three families of considerations that together provide strong support for the conclusion that our self-conception really is incompatibilist in at least some of its commitments. What I will now turn to is whether our self-conception is plausible.

## 4 Why not Libertarianism?

I have argued that commonsense is incompatibilist in its commitments, and, in particular, that it is libertarian. Why not embrace libertarianism, as Robert Kane argues we should? This answer can be put simply: there is no evidence that we are libertarian agents and a number of considerations that weigh against its likelihood.

Given the incompatibilist diagnosis of the prior section, a great deal might be taken to rest on whether we are libertarian free agents. For example, suppose for a moment that I am right that we have no evidence that we are agents of the sort described by libertarians. Now, imagine a criminal asking you to explain why he should be made to serve for a longer period of time, when the only answer we have appeals to a conception of agency *which has no evidence in its favor.* (Note: I am *not* saying that there are never good reasons for incarceration or denying various social services to different groups of people. Instead, I am asking you to imagine a case where the only reason for denials of service or parole is something like "You deserve to suffer because you freely (in the libertarian sense) made choices for which you must now pay a price.") It is the wrong kind of faith to suppose that the moral acceptability of our denying clearly valuable things to these agents is justified or even justifiable. It matters whether there is evidence that suggests we are libertarian agents, and so we must do better than to believe it on faith.

First, though, I need to make the case that libertarianism really does suffer from a lack of positive evidence and from various considerations that weigh against its plausibility, and, ultimately, its practical necessity. Sometimes philosophers talk about libertarianism as though it were obviously false, or as if the position were incoherent or self-contradictory. Perhaps it is, but I am

inclined to think that the best going accounts of libertarianism are none of these things. My dissatisfaction with libertarianism turns on two things. First, I am inclined to reject libertarianism on grounds of its comparative implausibility. Second, what empirical evidence there is that is relevant to the assessment of libertarianism does not favor the most plausible versions of libertarianism.

### *Empirical worries about libertarianism*

Just as there is no single libertarian view, there is no single empirical worry about libertarianism. The particular empirical concerns raised by libertarianism vary depending on the details of the libertarian theory. That said, there are empirical concerns that apply both to particular libertarian theories and to libertarianism as a whole. I will begin with some general empirical concerns about libertarianism as a whole before moving to particular concerns about the picture of libertarianism offered by Robert Kane in the first chapter.

The general worry is somewhat related to what Kane calls "the second prong" of the modern attack on libertarian free will (see chapter 1). The second prong holds that libertarian free will is impossible or unintelligible. I am not convinced of either, but I do doubt libertarianism's *plausibility*. My general concern can be stated simply: when contrasted with nearly any other view, libertarianism will turn out to be comparatively less empirically plausible than the alternatives. And, this is true even when we grant the possibility and intelligibility of libertarianism.

Here's why I think libertarianism is comparatively implausible: libertarianism requires that indeterminism be present in our agency in a very particular way, at very particular times, in the process leading up to or in the decision about what to do. Just how the indeterminism operates varies by the particular theory, but all libertarian theories are committed to indeterminism showing up in the world at particular times and places. (I am largely ignoring the possibility of an uncaused event but the point applies to libertarian theories that appeal to uncaused events.) In contrast, the alternatives do not have this requirement. Whether your favorite theory is a traditional form of compatibilism or the revisionist account I offer later in this chapter, these alternatives do not have this requirement. Moreover, what requirements the non-skeptical alternatives to libertarianism (i.e., compatibilism and revisionism) have will typically be requirements that libertarians have no special reason to dismiss. That is, libertarians typically do not deny that there is a range of conditions on free will and responsibility that are accurately rendered by compatibilists. (There is, of course, some disagreement between compatibilists about what those conditions are, precisely, but this is immaterial for present purposes.)

However, what marks out the difference between libertarians and compatibilists is not (typically) a dispute about there being compatibilist conditions on free will and moral responsibility. Rather, the dispute is over whether a further, indeterministic condition must be satisfied. So, for virtually any libertarian theory, there is an additional requirement it is bound to have, above and beyond the non-skeptical competitors.

Note that the point is not that we lack some special reason to think the world is indeterministic. Most parties could agree that there is good reason to think that at least some parts of the universe are indeterministic. The special burden of libertarianism is that it must hold that the indeterminism show up at particular times and places. Libertarianism is in this sense more demanding than non-skeptical alternatives. We do not know where science will take us and it takes a puzzling sort of confidence to simply assume that future discoveries will vindicate the more demanding theory. It might and it might not; it is an empirical issue that will be decided by the facts. For us, the issue is what we have reason to think will be more likely. Here, the history of speculative metaphysics should serve as a sobering reminder. By and large, the growth of human knowledge has not been kind to the products of the philosophical imagination. All things being equal, it seems a bad idea to bet on the truth of the more demanding theory.

The upshot of these remarks is that libertarianism faces a general worry about its plausibility: compared to any alternative, and in the absence of any evidence for the theory, these accounts will be less likely to be vindicated by future discoveries about the nature of human beings, all other things equal. The libertarian might argue that all other things are not equal. The libertarian could argue that there is some special reason why we should be committed to a picture of human agency with indeterminism nested in particular places and times along the pathway to human decisions, even if this makes the theory less plausible than alternatives. I discuss this possibility later but for the moment let us concede that the general plausibility worry only has force – to the extent that it does have force – if all other considerations are equal.

I claimed that there are two kinds of worries that libertarianism raises: one is a general worry about its comparative plausibility, the second are more specific worries tied to particular proposals on how to make sense of a libertarian conception of agency. The second class of worries varies by the particular libertarian account under consideration. I am inclined to think that there are serious plausibility worries raised by any worked-out account of libertarianism, but for present purposes I am going to focus on Kane's deservedly influential account of libertarianism to illustrate some of the specific worries about plausibility that can be raised against a libertarian account.

On Kane's account, paradigmatic instances of free will, what he calls SFAs, or "self-forming actions," are results of a particular kind of

indeterministic brain process. The idea is that in moments of conflict or uncertainty, when there are multiple but mutually exclusive aims we would like to attain, this stirs up a chaotic system in the brain that becomes sensitive to lower-level indeterminacies in the brain. (As Kane himself notes, chaotic systems are usually understood to be deterministic.) These low-level indeterminacies (presumably at the quantum level) influence an agent's decision by affecting the sensitive chaotic system generated by the agent's desiring mutually exclusive aims. The result is a SFA, or an instance of free will.

In connection with the prior criticisms about comparative empirical plausibility, it is worth briefly considering just how demanding the theory's commitments are: not only do agent mental processes have to turn out to be indeterministic, but they must also be indeterministic in a very particular way. If multiple mutually exclusive aims did not cause the brain to go into a chaotic state the theory would be disproved. If it tuned out that neurological systems weren't sensitive to quantum indeterminacies the theory would be disproved. If it turned out that neurological systems were sensitive to quantum indeterminacies, but not sufficiently sensitive to amplify quantum indeterminacies in a way that affects the outcome of choice, this too would disprove the theory. These are not marginal or insubstantial bets about what brain science will reveal to us.

Even if we interpreted it as a "just so" story – an account of one way we might satisfy the demands of our self-conception that does not conflict with the available data even if there is no positive evidence for it – it is hard to see why this would improve the situation by much. Mere possibility should strike us as an unsubstantial basis on which to our systems of praise, blame, and punishment. However, if we could do all the work that Kane's theory endeavors to do without these commitments we would have good reason to favor a less demanding theory. Later I will argue that we can do most, if not all of the work without appealing to his picture of free will. For the moment, though, let us assume that we do not yet have reason to suppose that we can do the work of libertarianism without indeterminism. Is there any reason to think that Kane's view has its own, independent plausibility?

As Kane notes, the existing literature on brain science does dovetail with some aspects of his account. The idea that the brain, or parts of the brain, might be chaotic under some conditions has been explored by some scientists. The main problem concerning the empirical plausibility of the view is that *there are no accepted scientific models of indeterministic brain events*. Indeed, virtually all brain science proceeds on the supposition of thoroughly deterministic explanations of the operations of the mind, and what evidence there is about indeterminacies in the brain weighs against indeterministic interpretations of brain phenomena. Moreover, what proposals there have been to locate a space for indeterminism-amplifying aspects of the brain (Penrose and

Hameroff's, for example) have been widely rejected by neuroscientists, philosophers, and mathematicians. Although brain science is by no means a complete and settled science it is clear that there are no widely accepted indeterministic models of brain activity, nor, for that matter, even an influential but contested model of indeterministic brain activity. In the words of Henrik Walter, a neuroscientist and philosopher, "to date there is no solid empirical evidence that local quantum phenomena play a role in neurons, and that there are good arguments to the contrary" (Walter, 2001, p. 162).

A proponent of Kane's view or one that similarly relies on the idea of brain-level amplification of lower-level indeterminacies might reply that all these considerations do is reveal a shortcoming of contemporary brain science. If brain science were properly pursued we might discover all sorts of indeterminacies in the head, perhaps exactly as Kane has described. Note that this sort of reply would essentially abandon the idea that there is some independent plausibility for a Kane-like view, by retreating to either a "just-so" story of the sort I criticized above or by arguing that brain science should take its cue from philosophy. One might justifiably complain that brain scientists are typically working with extremely impoverished conceptual models of human agency. So, the complaint might go, if they were working with a suitably sophisticated conception of human agency (such as Kane's) they might well interpret the available data and evidence in a way more favorable to indeterministic models of the brain.

I am less sanguine about the prospects for philosophy overturning brain science. It requires considerable optimism to think that armchair philosophy will be equal to or better than empirical brain science when it comes to revealing the structure of the brain. Nevertheless, I agree that much of contemporary brain science operates with a sometimes startlingly simple picture of human agency. And I agree that brain science could use some increased sophistication about philosophical categories and distinctions concerning human agency. What is notable, though, is that if there is a view about free will that brain scientists typically fail to see it isn't libertarianism. The nature and trajectory of what work there is on these issues typically bemoans (or celebrates) the implication that because the brain appears to operate deterministically there is no room for free will. Thus, the philosophical view of agency that is most frequently invisible to brain scientists is compatibilism, not incompatibilism. Indeed, if my prior diagnosis of folk incompatibilist commitments is correct, this is precisely what we should expect to find: brain scientists operating with latent assumptions of libertarianism, shifting to skepticism about free will when their models of the brain don't seem to require or make room for indeterminism.

Suppose I am wrong. Suppose neuroscientists and others simply have not been looking in the right ways or interpreting data in light of a libertarian

model of agency. Would this be of benefit to the libertarian fending off the charge of empirical implausibility? It is not clear that it would, at least not by very much. At best it would offer the libertarian the comparatively dim hope that were the sciences of the brain to be better informed by philosophy they might find what the philosophers were hoping for. It would not change the simple fact that there still is absolutely no evidence to suggest that libertarianism is true. And it would not change the fact that armchair philosophical speculation about the construction of the brain has an uninspiring track record. Any way you look at it, libertarianism's plausibility as a description of the form our agency in fact takes is clearly undermotivated.

I have focused on issues of comparative and empirical plausibility. As with nearly anything in philosophy, there is more than can be said for and against libertarianism. Instead, however, I want to turn towards considering what follows if we accept that we are not likely the sorts of agents described by libertarian theories.

## 5 Why not Hard Incompatibilism?

Thus far I have argued for two claims: (1) commonsense thinking about free will and moral responsibility have incompatibilist elements to them, and (2) it is implausible that we are libertarian agents. This might sound like a recipe for hard incompatibilism. It isn't. To be sure, skeptics about free will oftentimes arrive at their skepticism by just this sort of route. They offer some purportedly intuitive characterization of our agency and then they go on to argue that, as a matter of fact, we don't satisfy this picture of agency. As best as I can make out, though, these arguments simply do not work. Or, at any rate, they need to be supplemented by arguments that no one ever seems to provide. Let me explain.

Consider the kind of argument that Derk Pereboom offers in chapter 3. The argument starts by appealing to our judgments about what's intuitively a case of responsibility (or non-responsibility). From there, it proceeds to build an argument for hard incompatibilism. But why should we assume that our intuitions about free will tells us anything about the nature of free will? And, even if our thinking about free will was somehow a reliable guide to what truths there are about free will, why think that we cannot change our thinking (or at least out theorizing) about free will in some way so as to render it less problematic?

Recall the examples I began with in this chapter. In some earlier period, a careful study of our concepts of water, marriage, and magicians might have revealed a range of theoretical commitments in our concepts (and/or their conditions of application, if you favor this distinction). Eventually, it became

clear that the relevant part of the world wasn't cooperating with what pre-theoretically seemed perfectly intuitive. The stuff we called water wasn't really indivisible. Sometimes the relevant non-cooperating part of the world wasn't in nature but was simply us. Sometimes people called things marriages even when the involved woman wasn't thought of as property, and somewhere along the way we began paying magicians to entertain us without any expectation that they would demonstrate their mastery of supernatural powers. In all of these cases there was a transition period when there was not a neat mesh between what we referred to (the thing) and what we believed (the concept).

What the arguments of hard incompatibilists never do, as far as I can tell, is show that there could not be a similar disparity between our theoretical suppositions about free will and the nature of free will itself. As long as there is this gap in the argument, we are not entitled to conclude that the implausibility of our self-conception is evidence that we are not free and responsible, for we might have free will but it might be different than we tend to suppose.

I do not mean to rule out the possibility that our concepts of free will and moral responsibility really do settle the facts about these things, all by themselves. What I am pointing out is that we do not have any argument for supposing that they do.

Still, suppose that hard incompatibilists did offer us an argument of this sort. Even so, it would still not be clear that this would settle the issue in favor of the hard incompatibilist. There would remain two further challenges to embracing hard incompatibilism. First, the hard incompatibilist would need to show that we could not change the facts about free will by reconceiving how we think about free will. Hard incompatibilism would not follow even if we accept that our concept of free will settles the nature of free will for we might decide to change our concept. Second, even if changing the nature of free will by stipulation were impossible, we would still require some reason to not just call "free will" anything that satisfies the freedom-relevant capacities required for praise and blame. If there is a justification for our responsibility characteristic practices and attitudes, apart from libertarianism, then why not think that the freedom-relevant capacity that arises in those practices should count as a good re-anchoring for our usage and understanding of free will? Here too, some argument blocking this possibility is required.

I will discuss these two possibilities in order, beginning with the idea that we might revise by stipulation or fiat. Consider the nature of a touchdown. A touchdown is whatever it is that we (or some relevant subset of our community, anyway) say a touchdown is. Right now, a touchdown is 6 points in the context of a game of American football. We (or, again, the relevant subset of us) could change that. If the rules committee of the NFL decided to make

a touchdown worth 7 points or 5 points, then a touchdown would effectively become worth that new amount. Similarly, even if we accept that there is some tight connection between how we think about free will and moral responsibility and the nature of these things, we would still need some argument for thinking that we couldn't change those things. There are some apparent disanalogies between touchdowns and free will, and it might strike us as dubious that free will is as completely stipulative as touchdowns are, but this is something that merits discussion and something whose answer may depend on one's views about the nature of moral terms and related notions (if we accept that free will is a condition on moral responsibility). For example, if one were a conventionalist about moral terms, holding that the nature of moral terms are constituted or rest on complicated facts about human conventions, free will might more plausibly take on some of the stipulative character had by touchdowns. I do not mean to suggest that this must be so, or to argue for the conclusion that free will is subject to stipulation in as straightforward a way as touchdowns are. Rather, my point is simply that, again, the skeptic about free will has not offered an argument to block this possibility and there is some reason to think there may be a live philosophical possibility here.

Suppose we did receive such an argument. Here the second, further possibility I mentioned becomes live. Whatever else they do, our concepts of free will and moral responsibility are important for helping us to organize, track, and justify different ways of treating each other. If we can show that there is something that plays these roles, that does the work that is supposed to be done by these concepts, this may be a good reason to believe that free will and moral responsibility exist, even if they are somewhat different than we tend to have thought. At the very worst, it would show that there is something that is functionally equivalent (or nearly so) to free will and moral responsibility. The hard incompatibilist would need to show why an account of that nearby thing should *not* serve as an adequate replacement concept. If such an account would serve as an adequate replacement, then it seems entirely sensible to use 'free will' to refer to the freedom-relevant notion of power or capacity whose presence or absence is relevant to assessments of praise and blame. The hard incompatibilist might claim that such freedom is not *really* free will and that such practices are not *really* practices of responsibility, praise, and blame. Perhaps. But if these things do most or even all of the work that was to be done by our "real" concepts and practices, you might reasonably start to wonder whether the hard incompatibilist had the right account of what was "real" and if he did, whether there is any reason to go on caring about the "real" thing in the face of perfectly functional concepts, practices, and attitudes.

It is notable that, according to hard incompatibilists, free will and moral responsibility don't even really exist – in contrast to the purportedly "not real"

but very much existing and justifiable concepts, practices, and attitudes that we are imagining one might propose. Still, this observation only gains traction against the hard incompatibilist if we can indeed show that there is a justification for the responsibility characteristic practices, attitudes, and beliefs. So, a good deal turns on whether it can be provided. I will argue for this in a bit. What I have attempted to show in this section is only that an incompatibilist diagnosis of commonsense together with doubts about libertarianism does not, without further argument, warrant the adoption of hard incompatibilism.

## 6 Prelude to a Prescription: What Does the Indeterminism do, Anyway?

Before explaining how our responsibility-characteristic judgments, practices, and attitudes can be justified without appealing to libertarianism, I want to briefly remark on the purported importance of indeterminism.

Consider the question of how we go from being unfree agents to free agents. This is a puzzle faced by all accounts of responsibility, but there is something pressing about it in the case of libertarianism. As children we either had the indeterministic structures favored by your favorite version of libertarianism or we lacked them. If we lacked them as children, we might wonder how we came to get those structures. We might also wonder what the evidence is for thinking that we do develop said structures. Suppose the libertarian offers us an answer to these questions, and the other empirical challenges I raised in the prior section. We would still face another puzzle. What, exactly, does the indeterminism add? What follows in this section is not so much a metaphysical concern as it is a normative concern. It is a concern about what work the indeterminism does in libertarianism, apart from providing a way to preserve our default self-image as deliberators with genuine, metaphysically robust alternative possibilities.

Reflect for a moment on the connection between control and free will. Presumably, an act of free will is partly constituted by the agent having some control over what he or she does. You could hold that you act freely when you lack control, but this is an unattractive picture for free will when it is understood as the freedom condition on moral responsibility. An absence of control hardly seems like the way to become a responsible agent. So, responsible agency is going to presume some notion of control. Our question is this: what, precisely, is the work of the indeterminism? Is indeterminism required for control or is it required to elevate an agent that already has control into a free agent?

Suppose that the work of the indeterminism is to bestow control. Consider, however, the nature of an agent's first moment of free will. Inevitably, that moment will not derive from prior free aspects of character, inclination, standing policies, and so on. It is, after all, the *first* free act. What then makes it count as free, as the kind of thing that could underwrite attributions of responsibility? Presumably, the causal forces that lead to that first willing will be constituted by a web of events, inclinations, character traits, decisions, and so on over which the agent had no control. Out of these things a first free act is generated.

Now suppose the libertarian accepted that we could have responsibility-supporting control generated out of mental elements, the possession and nature of each of which was beyond our control. This strikes me as the kind of thing that we ought to say. There is disagreement about this, however. Some skeptics about responsibility insist that one must have control over all the elements that led to a free choice, and that this is impossible. However, recall the discussion in the prior section. Even if you thought that it was intuitively plausible that we must have control over all the elements that led to a free choice, this doesn't show we didn't have responsibility-supporting control: such control might be different than we imagined, or we might be able to change the facts by fiat, or there might be a suitably similar and fully workable notion of control that can serve as a replacement for the non-existing "real" control the skeptic says we lack.

So, if we do acknowledge that control can be attained out of elements that are not themselves controlled in their acquisition or content (as Kane acknowledges in chapter 1), it seems to me that we begin to sap some of the motivation for libertarianism as a prescriptive view. Here's why: it is hard to see what indeterminism adds to control, *given that the options indeterministically available to the agent were all products of things beyond the control of the agent*. In that first instance of free will, and in every instance that follows, what control the agent has is a function of what options the world bestowed on that agent (through experience, heredity, socialization, circumstantial luck, and so on). Any control the agent has must be built up out of those constraints. Given that even the indeterministic options are thus constrained, and the elements that gave rise to those options (experience, heredity, socialization, the circumstances one finds oneself in) were not in control of the agent, what does the indeterminism give the agent in the way of control? Why doesn't the indeterminism simply open up multiple paths to an agent, where the constitution and sources of those paths were not something over which the agent had control?

Now consider the alternative, where indeterminism doesn't bestow control but rather adds freedom to an agent that already has control. (Recall that on

the view Kane defends in chapter 1, indeterminism is actually a hindrance to control.) On such a picture, how could we make sense of control? Presumably, an agent could be said to have control by possessing some complex arrangement of agency, given a particular environment or range of environments. For example, control in an environment presumably relies on capacities to be sensitive or responsive to stimuli in the environment, the capacity to make decisions, the ability to reliably predict what effects one's actions will have on the environment and vice-versa. And, plausibly, none of these things requires indeterminism. (Indeed, this would seem plausible even if we weren't assuming that control does not require indeterminism). Indeterminism, then, is something superadded to control, something that transforms an already controlled agent into an agent with free will. This seems to preserve an important theoretical burden for indeterminism: it is the difference-maker between free and unfree action.

However, consider an agent that had all of the requisite capacities for control but lacked the indeterminism. Call him Max. The libertarian would insist that Max would not satisfy the freedom condition on moral responsibility. But what exactly would the freedom given by indeterminism provide for Max? It would *not* provide an additional measure of control – this possibility was ruled out above. Indeed, we might imagine that Max has all the control that anyone can have. If so, it is exceedingly difficult to see what indeterminism adds to maximal control. We might even put things this way: Max has all the control required for moral responsibility. Like anyone reading this book, Max deliberates about what to do, decides some things are better and some worse, and decides to do some things rather than others. The only thing he is lacking is indeterminism. Were he to suddenly be bestowed with it (in whatever way the libertarian likes), this wouldn't change the way his deliberations appear to him. He would still be deciding between options. He would still (let us say) have just as much control as he had previously. And the mental elements out of which his control was constituted and out of which the indeterministic possibilities would be shaped would not suddenly become under his control if they were not already. So, whatever freedom it bestows on Max it is nothing that changes the way his deliberations will appear to him and it does nothing to change the control that he actually has. The work indeterminism does begins to seem ephemeral.

The libertarian might be tempted to reply that the work left over for indeterminism is crucial in at least the following respect: without it, Max would fail to be an intuitively free and responsible agent. This can be an important consideration, and I ultimately agree about how the intuitions sort out. But we are beyond mere intuition description. We can grant that commonsense has these commitments. What we are not trying to do is to determine if there is anything besides our self-image that hinges on the success or

failure or our turning out to be indeterministic. What we need is an explanation of what *normative* work indeterminism does in generating responsibility. It is difficult to see what explanation the libertarian might offer.

Perhaps the libertarian will be tempted to respond with a burden of proof argument. That is, the libertarian might argue that although he or she is inclined to think that indeterminism is required for free will, and thus responsible agency and the integrity of our responsibility practices, critics of libertarianism are not really any better off. Those who think you can get responsibility without indeterminism have not successfully shown their case either, as we lack an adequate defense of how you could have genuinely responsible agency (and correlatively, responsibility practices) in a deterministic world. Until such an account is in hand, the libertarian has no more burden to prove the truth of his view than a critic has a burden to prove the truth of her view.

I am dubious about whether the argumentative burdens are really so equal as this line of response suggests. However, we might simply proceed to showing that we can make sense of responsible agency and justify our practices of responsibility without appealing to indeterminism. This is exactly what I will turn to now.

## 7 Prescription: An Outline of a Moderately Revisionist Approach to Free Will

Thus far I have argued that the shortcomings of commonsense conceptions of free will and moral responsibility do not, by themselves, warrant hard incompatibilism. Moreover, I have argued that it is very unclear what work, apart from protecting our commonsense conception of ourselves, the postulation of indeterminism is supposed to provide. What remains to be shown is how we might justify the web of practices, attitudes, and beliefs that are characteristic of our attributions of free will and moral responsibility. If we can show how this might be done, then we will have a revisionist account of free will and moral responsibility along with some principled reason to reject hard incompatibilism as the right response to the troubles of libertarianism.

Although I will proceed to describe my version of revisionism, it bears noting that there is an enormous range of theories that might legitimately be called revisionist. A theory counts as revisionist in the sense I am interested in if it offers a different prescriptive theory of responsibility (an account of what we ought to believe) than it offers for a diagnostic theory of responsibility (an account of what we tend to believe). However, this characterization of revisionism permits a number of possibilities. One might hold, for instance, that commonsense is compatibilist, but that we ought to believe an

incompatibilist conception of free will. This would be a kind of revisionist incompatibilism. And, one can imagine a variety of different diagnoses about common sense (e.g., that it is committed to agent causation, or committed to uncaused events), where the prescription is similar in terms of the compatibilism/incompatibilism issue, but nevertheless different in its particular formation (perhaps event causal libertarian is what we should revise in the direction of). None of these are the sorts of view I favor.

I have been arguing that there is good reason to think that an accurate diagnosis of commonsense will acknowledge the presence of incompatibilist elements in our thinking (minimally, metaphysically robust alternative possibilities). And, for some of the reasons I have presented, I doubt that we can make good on those elements. So, in broad terms, the revisionist proposal I am offering is *a hybrid account: incompatibilism about the diagnosis and compatibilism about the prescription*. Alternately, we might say the account is incompatibilist about the folk concept of free will and compatibilist about what philosophical account we ought to have of free will.

There are two ways we might pursue a revisionist justification of our responsibility-characteristic practices, attitudes, and beliefs. One route we might call *revisionism on the cheap*. We can call the other route *systematic revisionism*.

Revisionism on the cheap is gotten by taking your favorite compatibilist proposal of free will and declaring that it is not beholden to commonsense intuitions about responsible agency and free will. Instead, the positive compatibilist proposal is a purely prescriptive account of how we ought to think of free will and moral responsibility. The virtue of this approach is twofold: it is both methodologically simple and endowed with the not insubstantial resources of already existing compatibilist theories. The shortcoming of this approach is that it relies on existing compatibilist theories, which are largely generated under non-revisionist constraints. Most contemporary compatibilist theories were not developed under an explicitly revisionist conception. As such, there will likely be elements of any such theory that are attempts to accommodate some aspect of commonsense that a revisionist account need not accommodate.

Considerations such as these may raise the worry that revisionist accounts are not beholden to anything, a moving target on a shifting philosophical landscape. One might even wonder if revisionist accounts are vacuously immune to all counterexamples simply in virtue of declaring that they are not attempts to capture commonsense. Although this is a natural worry to have about nearly any revisionist proposal in any domain, it is not applicable here. First of all, the revisionist about free will is not entitled to be revisionist about any aspect of the theory that is inconvenient. At least initially, the only thing the revisionist about free will is justified in revising is anything that both

proceeds from the difficulties that are embedded in or derived from the problematic commonsense notions of freedom and responsible agency and that can be repaired without appealing to those troubled notions. Second, as with any proposal, a revisionist account will be constrained by considerations of consistency and coherence. Contradictory claims or implications will be out of bounds here. Finally, a revisionist account will be constrained by two particular standards: a standard of naturalistic plausibility and a standard of normative adequacy. The standard of naturalistic plausibility demands that any proposal be compatible with a broadly scientific worldview. The standard of normative adequacy requires that the prescriptive theory of free will function appropriately with respect to the various normative burdens of a theory of free will.

The alternative to revisionism on the cheap is systematic revisionism. Systematic revisionism must obey the constraints described above, including standards of naturalistic plausibility and normative adequacy. What makes a systematic revisionist distinctive is that it proceeds from the ground up on the basis of attempting to provide an intentionally and explicitly revisionist proposal of free will. Unlike a compatibilist proposal recast as revisionism on the cheap, a systematically revisionist account is conceived of from the start as an intentionally revisionist account. It straightforwardly acknowledges the incompatibilist elements in commonsense thinking and proceeds on the basis of difficulties in satisfying the theoretical commitments of commonsense.

When building a systematically revisionist account of free will, it helps to be clear about exactly what we aim to provide with an account of free will. As I noted at the beginning, I am following the bulk of the literature, both current and historical, in supposing that free will is the freedom condition for moral responsibility. Thus, if we want to understand the nature of free will, we would do well to understand the nature of moral responsibility and what role a freedom condition might play on responsible agency. (As an aside: thinking about free will in this way does raise a puzzle about semicompatibilism: how is it different than more traditional forms of compatibilism? What does the "semi" add, given that it holds that responsibility is compatible with determinism and given that it contains an account of the freedom-relevant condition on moral responsibility, i.e., free will? Is there a reason ordinary compatibilists cannot similarly concede that there is at least some sense (not relevant to responsibility) of the ability to do otherwise that is incompatible with determinism being true?)

The most useful initial characterization of the conceptual role for moral responsibility is as something that plays an important role in our organization, coordination, and justification of differential treatment of one another. In particular, it is connected to praiseworthiness and blameworthiness. In turn,

judgments of praiseworthiness and blameworthiness underwrite a web of emotional reactions, judgments, and social practices that can include (but are not limited to) reward and punishment. A responsible agent is thus the kind of agent that can be appropriately judged as responsible or not. The details of whether or not a particular responsible agent can be praised or blame for a particular action turn on facts about both the agent and the norms of responsibility. Whether an action deserves praise or blame depends on (1) whether the agent that did it is appropriately subject to norms of responsibility (young infants are presumably not, normal mature adults presumably are), and (2) what the norms of praise and blame say about actions of that sort in the relevant sort of context.

A satisfactory revisionist account of free will will therefore be informed by two different accounts of how distinct conceptual roles might be satisfied: (1) an account of responsible agency and (2) an account of the responsibility norms. With respect to the latter, our present concern is not so much with the particular content of the norms of responsibility. Instead, we need to know about the nature and justification of the responsibility norms so as to inform our account of the freedom condition on responsible agency. So, the sort of account of the responsibility norms that would be most useful is not some specification of the particular norms that we face but rather an account of the general structure, aim, and source of justification for the responsibility norms. In what follows I offer a sketch of both the nature of responsible agency and the nature of the responsibility norms.

## 8 Justifying Praise and Blame Without Libertarianism

Taken as a whole, the responsibility norms and their attendant social practices, characteristic attitudes, and paradigmatic judgments constitute what we can call *the responsibility system*. The challenge for a systematic revisionist about free will is to tell a story about the justification of the responsibility system, and, in particular, the responsibility norms, that will give us some principled grounds on which to offer a naturalistically plausible and normatively adequate account of moral responsibility.

The details of this sort of account will depend, in part, on the details of our conception of responsible agency. However, a theory of responsible agency will need to be integrated into some broader account of the point of a system of responsibility. So we will need interlocking accounts of both the nature of responsible agency and the nature of the norms of responsibility. I'll begin with the former.

The importance of being a reasons-trading creature has been a mark of a wide range of philosophical accounts of agency and morality since at least

Plato. Within the narrower limits of recent theories of responsible agency the role of reasons has loomed large. On the particular account I favor, the distinctive mark of the freedom-relevant aspects of responsible agency is the agent's sensitivity to specifically moral considerations and the capacity of that agent to appropriately govern his or her conduct in light of those considerations.

As I am using the terms, considerations are, roughly, the kinds of things that can generate reasons. Moral considerations are considerations with moral significance, and, as such, are the kinds of things that typically work to generate reasons for action against a background of beliefs, moral norms, agent values, and perhaps agent motivations. Thus, on this account, what makes us appropriate targets of the distinctively moral form of evaluation that is the hallmark of responsibility assessments is our ability to detect moral considerations and appropriately guide on conduct in light of what reasons they generate.

As I noted, this basic idea is not novel to this account and it has a wide range of able proponents (for an influential version, see John Martin Fischer's account outlined in chapter 2). What is distinctive about my account is the justificatory story it is taken to generate. As I see it, carefully reflecting on the importance of moral considerations for responsible agency yields a simple, but powerful idea about the aim of system of responsibility: the responsibility system aims to get creatures like us to better attend to what moral consideration there are and to appropriately govern our conduct in light of what moral reasons those considerations generate. Assessments of praiseworthiness and blameworthiness are not *merely* reactions we happen to have to one another (although they are partly that). They play a special role in getting us to be better beings, agents better attuned and more appropriately responsive to moral considerations and the reasons they generate. To use some lofty language for a moment, what the responsibility system does is to foster the flourishing of an intrinsically valuable form of agency. When you judge me blameworthy for being insensitive to someone's feelings, the sting of your disapproval forces me to attend to considerations that I might have failed to see or failed to act on in the right way. Over time, and given widespread participation in this system of judgments, practices, and attitudes we come to help both ourselves and other consideration-sensitive creatures to better track what moral considerations there are.

These ideas are simple. What makes them important is that they provide a powerful framework for explaining how our responsibility-characteristic judgments, practices, and attitudes can be justified without appealing to the libertarian elements latent in commonsense. The responsibility-characteristic practices, attitudes, and judgments are justified inasmuch as they, on the whole and over time, tend to contribute to our better perceiving and

appropriately responding to moral considerations. It is plausible to think that our responsibility-characteristic practices, attitudes, and judgments have this effect. Think about moral praise and blame: they tend to get creatures like us to pay better attention to the moral considerations recognized in our sociohistorical context and they provide incentive for us to act accordingly. Since they are reasonably effective at doing this, it is plausible to think that the responsibility system is by and large justified. To be sure, there are likely aspects of our current norms that are less than fully ideal. It would be unduly optimistic to assume that the exact norms that have the most currency in our society just happen to be the normatively ideal norms. Nevertheless, it is plausible to think that the bulk of our responsibility-characteristic practices, attitudes, and beliefs can be justified in this way.

That we can justify the bulk of our responsibility-characteristic practices and attitudes in this way does not preclude other ways of justifying the responsibility system or parts of it. Perhaps there are overlapping justifications available, and, if so, this will raise some interesting further issues. The relevant issue for us is whether the responsibility-characteristic practices can be justified without appealing to some picture of libertarian agency. It should be clear that they can be: the sort of account I sketched is one that made no appeal to libertarianism and simply relies on the ideas that (1) we are creatures who can, at least sometimes, detect moral considerations and appropriately guide our conduct in light of them, and (2) the responsibility system as whole and over time tends to get creatures like us to detect and appropriately respond to moral considerations.

(I want to flag that one upshot of the view I propose is that the responsibility-relevant notion of "can" to which free will attributions are tied will not be a libertarian one. This is not to say that we will do without a notion of "can" – it is merely to note that the relevant notion of can will be construed differently than we might intuitively suppose. So, some notion of "ought implies can" will be preserved, but the operating notion of can will not be libertarian. More on this in the next section.)

One might protest that the notion of "being able to detect moral considerations" and the related notion of "appropriately responding to moral considerations" might sneak back into notions of libertarian agency. They need not. To see why, consider a very simple picture: suppose that all non-human action is deterministic. Suppose also that there is a range of considerations we can call "good treatment considerations" that for the right kinds of creatures generate reasons of a particular sort. Even if your dog were deterministic, we might train your dog to become sensitive to those good treatment considerations. When the dog treats a child well you shower it with approval and attention. When it treats a child poorly, you heap scorn and abuse on it. It is reasonable to think that over time the dog will get pretty good at

tracking good treatment considerations and governing his or her behavior in light of them. Perhaps this happens by habituation, conditioning, animal reasoning, or some combination of these things. The point is that it does happen. Now move to the case of humans. There are obviously important cognitive differences between humans and dogs. These differences might be relevant for being able to move from simple sensitivity to considerations to full-blown reasons-deliberating agency. It depends on your picture of reasons and whether other creatures can have reasons. I do not see any reason to suppose that only humans operate in the space of reasons, but nothing turns on this point. However, there is no reason to suppose that the *addition* of those cognitive elements that are the difference between humans and dogs would somehow remove the ability (and, by stipulation, the addition of indeterminism) that is already present in dogs. If anything, we should expect that increased cognitive complexity brings with it an increased range of considerations that one could be sensitive to and, perhaps, increased ability to appropriately move from considerations to reasons that appropriately influence deliberation and action in a robustly rational creature. These cognitive differences might further provide grounds for thinking there is something special or distinctive about human beings. Be that as it may, these cognitive differences do not provide a reason for thinking that sensitivity to considerations and the ability to appropriately govern one's behavior in light of them could only be had by humans in an indeterministic universe.

A different kind of concern about my account of the justification of the responsibility norms might be expressed by dissatisfaction at the account being "merely consequentialist." This strikes me as either wrongheaded or deeply misleading. First, it is not clear why, even if it were true, this should be an objection. Consequentialism is one of the major contender theories of morality, and inasmuch as we are discussing something plausibly in the domain of the moral one of the major theories in philosophical ethics ought to be a perfectly respectable view to hold. Nevertheless, this will surely fail to persuade anyone inclined to have raised the criticism in the first place. A more nuanced reminder might be more helpful: even if consequentialism isn't a complete moral theory, it is overwhelmingly plausible that consequences can be extremely important. Second, although the account I have offered does appeal to something that might fairly be called "consequential reasoning" it does not follow that the norms of responsibility are themselves consequential*ist* in the sense of being committed to a consequentialist theory of normative ethics. Indeed, there is no obvious reason why this account is not be compatible with a wide range of ethical theories, including Kantian theories (which are oftentimes taken to be the class of theories most opposed to the picture described by consequentialism). Kantian moral theory does not preclude one

from *ever* reasoning consequentially, although it does preclude consequences from playing a core role in some specific moral notions. This is one of those cases where the Kantian might recognize a minimal role for consequences. By participating in the responsibility system we participate in a practice that fosters a special kind of agency, reasons-mongering agency, in others and ourselves. Indeed, we might suppose that one important way to respect the rationality in oneself and others is to participate in a system that fosters it.

I have offered an account of the *justification* of the responsibility system. It is not an account of how we tend to think of our own judgments of praise and blame, nor of what, if anything, we think justifies these judgments while we are making them. We might never praise or blame with an eye towards influencing each other in the ways I have described. Even if we came to accept a revisionist account of moral responsibility it is doubtful that we would make judgments with an eye towards the effectiveness of particular or general judgments of responsibility. Instead, we would judge people as responsible in much the same way we do now – albeit without some libertarian commitments – by making assessments about the quality of their actions in light of the norms we accept.

It may be helpful to think about an analogy. The justification for making foul calls in a sport turns on the benefits the foul calls have for the players and the conduct of the game. However, when a given referee calls a foul it does not follow that he or she has in mind the benefits of foul calling for the game. Instead, he or she simply recognizes and decrees a violation of the relevant norm. Even on a revisionist account, particular assessments of responsibility will rarely if ever attend to justification for the practice. Instead, it will straightaway appeal to a norm of praise or blame. Nevertheless, what justifies those norms and attendant practices are its effects on getting us to attend to moral considerations.

Moreover, there is good reason for doubting that *the content* of the responsibility norms (as opposed to the whole system) will have anything like a simple consequentialist structure to them. The familiar difficulties of calculating exactly when and how much moral scorn will be maximally effective are relevant here. Instead of explicitly consequentialist responsibility norms we should instead expect that the justified system of responsibility norms will look very familiar, accommodating both backwards-looking attitudes (such as gratitude) and forward-looking attitudes and practices. What makes someone responsible in a particular instance is not settled by whether or not holding such a person responsible gets the best results with respect to fostering responsiveness to moral considerations. What makes you responsible is determined by whether one is the right sort of agent (i.e., one that is sensitive and capable of guiding conduct in light of moral considerations) in the relevant context and what the norms of responsibility say about agents in those

circumstances. In turn, these norms will be determined by what is conducive to the aim of the responsibility system over time as well as what is permitted and required by the True Theory of normative ethics (whatever that turns out to be).

For creatures like us, with psychologies brimming with a wide range of powerful attitudes and concerns but fairly limited powers of calculation about these same things, the most effective set of norms over time will permit and perhaps require less cognitively demanding norms such as "express dissatisfaction about normal adults who are insensitive to the feelings of others." Practicality matters for efficacy. Simple norms can quickly yield complex difficulties that are tough to untangle and so there will always be thorny questions to be sorted out by moral philosophers, lawyers, and students willing to engage in philosophy. (To see how simple norms can generate complex situations, just consider how relatively simple rules in games like Go or Othello yield immensely complex games.) What should be clear is that the justified responsibility norms and the kinds of attitudes, practices, and judgments they permit are not restricted to simple act consequentialist-style decision procedures and it is a mistake to suggest that they would be; compatibility with Kantian and other moral theories is surely the case.

The relatively general and formal structures of the responsibility system I have described depend on details that do not emerge simply by reflecting on the nature of responsibility by itself. An account of the content of the responsibility norms will appeal to some broader account of normative ethics to determine the content of the norms and to provide specification of what, precisely, constitutes a moral consideration. For example, when the justificatory story I sketched is incorporated into a broadly Kantian approach to ethics, various prohibitions will emerge that constrain the contents of the responsibility norms. For example, classic scapegoating worries, which are often raised as difficulties to consequentialism (such as killing an innocent man to appease a mob), cease to be worries on this account because scapegoating would be ruled out by the background Kantian commitments. This is because considerations about using people merely as means to an end are typically taken by Kantians to provide adequate justification for prohibiting scapegoating, even when it achieves some otherwise desirable result (such as increasing happiness). Presumably, core commitments of the ethical theory such as these would trump the kinds of things that one might otherwise expect on a theory that justified praise and blame in light of consequential reasoning. So, even if scapegoating increased sensitivity to moral considerations in a particular case (or even more generally), a system of moral responsibility embedded in a broadly Kantian normative ethical theory would not permit blaming an innocent person for some transgression he or she did not permit. Indeed, it may well be the case that a kind of regard for the

reasons-mongering agency of others is part of what makes responsibility practices *responsibility practices*, and not some other thing.

It is thus plausible to think that the bulk of the responsibility system can be justified independent of whether we have libertarian agency and in a way that will be compatible and well integrated with whatever our best normative ethical theory turns out to be. If so, then even if we accept that commonsense thinking about free will and moral responsibility is incompatibilist, and even if we accept that libertarianism is comparatively implausible, the move to hard incompatibilism remains unjustified. What is justified is a moderately revisionist repair to commonsense thinking, one that strips away the metaphysically demanding elements of libertarianism and preserves the justifiable core of our attitudes and practices.

## 9 What about Free Will?

Earlier I claimed that if we can show that there is something that does the work that is supposed to be done by the concepts of free will and moral responsibility, then this would provide us with a good reason to believe that free will and moral responsibility exist even if it is somewhat different than we might tend to have thought. This is just what I have attempted to show. Even in the worst-case scenario, where an incompatibilist might insist we do not "really" have free will, we would have a good candidate for a replacement concept in the freedom-relevant notion of power or capacity whose presence or absence is most salient to assessments of praise and blame.

What we need, then, is an account of the freedom-relevant condition on moral responsibility on the moderately revisionist account I am proposing. Here, we must look to the picture of responsible agency that motivated the justificatory picture I provided. Responsible agency has two principal requirements for responsibility-supporting freedom: a detection or sensitivity requirement and a self-governance condition. Where free will is to be located, then, is in the satisfaction of these conditions. An agent can be said to have free will or to be acting from or with free will when that agent, in the context of deliberation or action, has the capacity to detect moral considerations and can govern him or herself appropriate way in light of those moral considerations. The relevant sense of capacity is not one that requires indeterminism. Instead it is one that will (by design – for this is a revisionist account) be compatible with determinism and indeterminism.

One might wonder whether there are any good candidates for a sense of "can" compatible with determinism. The answer is clearly yes. To see why, consider the capacity to dance to merengue music. Suppose you have that capacity (it is a good one to have – you might consider acquiring it if you lack

it!). If you have it, it is surely a capacity you retain even when you are sleeping. You might lack the opportunity to exercise the capacity while sleeping, but it would be strange to say you *lost* it. Where would it have gone? How would you relearn it so quickly when you woke up or when the music started to play? Rather than saying you lose the capacity when asleep it makes more sense to say that you retain the capacity, but that you lack the opportunity to exercise it while sleeping. Specifying the *exact* nature of the capacity can be difficult and it is here, too. However, we could say something like "you have the capacity to dance as long as it is the case that were you in the right circumstances, you would (were you to be so disposed) dance in the characteristically merengue way."

Now turn to the case of determinism. The antecedent conditions of the universe (deterministic or not) surely play a role in whether or not you have the capacity to dance the merengue. For example, if things had been slightly different, there would have been no planet Earth, and thus, no merengue music or dance. And, the conditions of the universe might play some role in whether you continue to have the relevant capacity. Amputation of your legs would very likely impinge on your capacity to dance. However, the hypothetical fact of determinism would not, by itself, show that you lack the capacity to do merengue dancing. At most, what determinism would settle is whether it was determined that you should have an opportunity to exercise this sort of capacity, should you have it. But even if it was determined that you would never have the opportunity to dance merengue after your 30th birthday, it would still be the case that (barring injury of the relevant sort) you would continue to retain the capacity to dance for some time afterwards: it would simply be unexercised, like many of our capacities.

Here, some readers might suspect some philosophical trickery, a return of the wretched subterfuge so many non-incompatibilist accounts lapse into. It may be useful to re-emphasize that I am *not* saying that this is how we do tend to think of free will. On the contrary, I am happy to acknowledge that current commonsense tends to require something more, something like the robust alternative possibilities described in the Garden of Forking Paths model of freedom. This is true even if commonsense also requires the kinds of things I have specified. What I am claiming is that we *should* think about free will along the leaner, revised lines I have sketched. It has the significant benefit of doing the work we require without the disconcerting difficulties entailed by the commonsense picture.

I think it is fair to say that the philosophical consensus about these things recognizes that there are many legitimate senses of capacity whose appropriate use persists even if determinism were true. If the consensus is right, there is plenty of conceptual room to locate a construal of the revisionist-relevant sense of a capacity for self-governance that would persist even in the face of

determinism. Such a capacity need not be flawless. I might have the capacity to behave courteously, but it does not follow that I do so every time the opportunity arises. Sometimes I will non-culpably fail to recognize considerations of courtesy. Other times, however, I might simply suffer weakness of will and start nibbling on the food in front of me before the plates of my dinner companions arrive. So, we need not reject the possibility of weakness of will and we have plenty of senses of capacity that seem to allow for it or things that are analogous to it. To be sure, the precise details of the revisionist sense of capacity will be a difficult one to settle. Nevertheless, these considerations should make it clear that there are a number of viable options. To put the lesson simply: we can plausibly stipulate that an agent has free will when he or she has the capacity to detect moral considerations and can (in some broadly compatibilist-friendly way) appropriately guide his or her behavior in the right way.

There is always more philosophical work to be done. However, this completes the main outlines of the revisionist theory of free will and moral responsibility that I recommend. Still, some final epicycles may be worth mentioning.

First, I am skeptical that we can provide an account of any single mechanism or faculty involved in the detection of moral considerations and I am similarly skeptical about there being a single or general capacity for self-governance. Our capacity for self-governance (of the sort required for free will) is very context-specific, dependent on facts about our selves and the context we are in, oftentimes in ways that are invisible to us. The psychological mechanisms that provide us with what we are entitled to call "free will" in one context are oftentimes of little use in other contexts. My ability to steel myself effectively against a certain form of temptation (another shot of Don Tacho tequila) may evaporate with comparatively minor changes in context (such as my brother entering the room). Even on this revisionist account of it, free will remains a fragile achievement of a special, sophisticated kind of agency. We would therefore do well to acknowledge that that there will be cases where free will is absent where we currently tend to think it present. Less frequently (for I suspect we tend to overestimate the power of our agency, at least in the West), there will be cases where our free will is more present than we currently tend to think.

## 10 Recovering our Freedom

At least one important thread of commonsense thinking about free will and moral responsibility is incompatibilist. Any theory that fails to acknowledge this is going to end up seeming like what Kant called a "wretched subterfuge"

or what James called a "quagmire of evasion" – at least to a good many of us. What free will skeptics such as Derk Pereboom get right is the idea that we are in trouble if free will and moral responsibility depend on libertarianism. Contemporary versions of it, such as Robert Kane's, are hardly committed to "rape and perversion of logic" that Nietzsche seems to have ascribed to these views. Nonetheless, they face serious difficulties that we should not underestimate. What compatibilists (including semicompatibilists such as John Fischer) get right is that we can get along without a libertarian conception of agency. Revisionists, however, part ways with traditional compatibilists about whether commonsense thinking about free will and moral responsibility really is compatible with determinism. Instead, we should construe compatibilist pictures of agency as a replacement and upgrade of commonsense. However, we should not suppose that this replacement *is* commonsense, or is beholden to intuition tests grounded in our current error-plagued conception of agency. To do that is to invite the (justifiable) charge of evasion and subterfuge that have plagued compatibilism. Even semicompatibilism, which surely goes further to recognize the power of incompatibilist intuitions more than most compatibilist accounts, does not go far enough in recognizing that a naturalistically plausible and normatively adequate account of free will and moral responsibility will require abandonment of important parts of our self image. These are aspects of our self-image that are not merely the inventions of philosophers or extravagant demands detached from widely shared concepts. They are pervasive and deep aspects of our commonsense.

To put the point somewhat differently, the revisionist's construal of the equilibrium point for philosophical reflection on these issues is further away from commonsense than the (even the semi-) compatibilist will admit. This is not to say that we should dismiss the use of commonsense intuitions here or in any other domain of philosophy. Inasmuch as philosophy is concerned with issues where we lack reliable methods for determining what the truth is in some particular domain, linguistic and conceptual intuitions will surely have an appropriate role to play. Revisionism does not claim that we should dispense with intuitions altogether. Instead, it is a reminder that sometimes the equilibrium point between our pre-philosophical thoughts about an issue and what we learn by conceptual and empirical investigation is oftentimes some distance removed from our initial pre-philosophical position. Still, some amount of caution is in order. We do not want to prescribe surgery for every problem, as repair by removal is sometimes no repair at all.

Revisionism does face an important difficulty: a shift in our web of beliefs can be a threatening thing. It is all the more so when it concerns our image of ourselves. No one enjoys admitting that he or she is mistaken, and the reluctance can become insurmountable when the subject of error is oneself, or, even worse, some aspect of oneself that we take to be central and important

to our shared lives together. So, I think, it is not unusual to have lingering resistance to revisionism, even if one accepts most of the picture I have been sketching. This resistance may be even more pronounced in people whose libertarian intuitions depend on elements I have said comparatively little about (for example, agent causation and intuitions about ultimacy). Even for these views, though, the point remains the same: we can do the important work of these notions without needing to be agent causes or ultimate sources of our action. We need not refuse to acknowledge our incompatibilist intuitions, but we also need not assume that so much depends on them.

## Further Reading

There is a large literature on the traditional philosophical arguments for and against incompatibilism, some of which can be found in the further readings listed after the chapters on libertarianism and compatibilism in this volume. For some of the recent experimental work mentioned in this chapter, see Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner, "Is Incompatibilism Intuitive?," *Philosophy and Phenomenological Research* (forthcoming); Shaun Nichols, "Folk Intuitions on Free Will," *Journal of Cognition and Culture*, 6(1 & 2) (2006); Shaun Nichols and Joshua Knobe, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions" (forthcoming); Robert L. Woolfolk, John Doris, and John Darley, "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility," *Cognition* (2006).

For an in-depth and rigorous exploration of many difficulties facing libertarian views see Randolph Clarke's book, *Libertarian Accounts of Free Will* (Oxford: Oxford University Press, 2003). For an examination of some of the neuroscience issues, see *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy* (Cambridge, MA: MIT Press, 2001). Further elaboration of my concerns about libertarianism and hard incompatibilism can be found in "Libertarianism and Skepticism about Free Will: Some Arguments Against Both," *Philosophical Topics*, 32(1&2) (2004), 403–26.

Revisionist threads can be found throughout much of the existing literature on free will and moral responsibility. In spite of this, there have been comparatively few attempts to pursue it in an explicit and systematic way, especially when we distinguish the view from more straightforwardly skeptical views. J.J.C. Smart's "Free Will, Praise, and Blame," *Mind* 70 (1961), 291–306 is probably the place to start for contemporary work that is revisionist in aim, although Smart's construal of responsibility is somewhat idiosyncratic. Other places to look for statements or defenses of explicitly revisionist

views about free will and moral responsibility include Richard J. Arneson, "The Smart Theory of Moral Responsibility and Desert," in Serena Olsaretti (ed.), *Desert and Justice* (Oxford: Oxford University Press, 2003); Ira Singer, "Freedom and Revision," *Southwest Philosophy Review*, 18(2) (2002), 25–44; Manuel Vargas, "The Revisionist's Guide to Responsibility," *Philosophical Studies*, 125(3) (2005), 399–429; Manuel Vargas, "Responsibility and the Aims of Theory: Strawson and Revisionism," *Pacific Philosophical Quarterly* 85(2) (2004), 218–41; Henrik Walter, *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy* (Cambridge, MA: MIT Press, 2001).

For an interesting and prominent case of someone frequently interpreted (wrongly, I think) as revisionist, see Daniel Dennett's two books on free will (*Elbow Room* (Cambridge, MA: MIT Press, 1984) and *Freedom Evolves* (New York: Viking, 2003)). Jonathan Bennett's influential interpretation of P.F. Strawson's "Freedom and Resentment" has some clear revisionist sympathies (see his "Accountability," in Zak Van Straaten (ed.), *Philosophical Subjects* (New York: Clarendon, 1980)). Finally, Thomas Nagel's discussion of freedom in the context of internal and external standpoints explicitly recommends revisionism about what he terms "autonomy," although he is less sanguine about revisionist prospects for responsibility (see *The View From Nowhere* (New York: Oxford, 1986)).

# 5

# *Response to Fischer, Pereboom, and Vargas*

*Robert Kane*

## 1 Introduction: Forking Paths Again

In response to the other three contributors, let me begin on a personal note. If you are like me, you react to philosophical debates about free will in the following way. At first, things seem pretty clear. There seems to be some kind of conflict between free will and determinism. If every occurrence in the universe were antecedently determined by Fate or the decrees of God or the laws of logic, by the past and the laws of nature, or other factors beyond our control, there would seem to be no room for free will. But the more philosophy you read, the more complicated things get. After the philosophers have spun their webs and drawn more and more of their epicycles to support one view or the other, you tend to get increasingly confused and lose sight of what bothered you about free will in the first place. Now, don't get me wrong. The issues *are* complicated and we philosophers are just doing our thing. But it is easy to lose sight of the forest for the trees as the philosophical mist grows thicker; and you may have had that feeling as this volume proceeded.

So let us return to basics; and with regard to free will, this means returning to the "garden of forking paths" image cited at the beginning of my chapter 1.

Free will seems to require that at some points in our lives there are multiple pathways into the future that are "open" to us; and it is "up to us" which of these pathways we choose. Yet determinism implies that at any given time, there is only one possible path into the future, given the past and laws.

This looks like a plain contradiction; and it is one reason why libertarians like me believe that *indeterminism* is required for free will. Now compatibilists say that indeterminism isn't required for free will; we can have all the free will (and moral responsibility) worth wanting even in a determined world. I

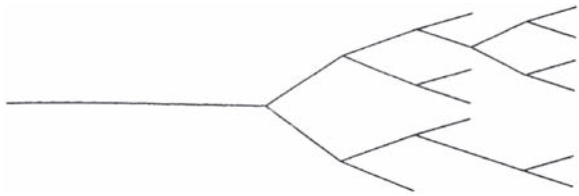

**Figure 1** Garden of Forking Paths

doubt this; but they have arguments that must be addressed. In addition, many critics of the libertarian view (including the three in this volume) also say that that indeterminism wouldn't *help* with free will, even if it did exist, because indeterminism would just amount to chance; and chance isn't free will either. I also disagree with this and have argued that indeterminism does not necessarily mean "mere chance." (We'll be returning to this point again.)

But these charges make it worth reminding ourselves *why* we libertarians think indeterminism is required for free will in the first place. The answer is that, while indeterminism need not mean mere chance, what it *does* mean is that there are "multiple paths into the future we might choose, given the past," as the garden of forking paths requires, whereas determinism means the opposite, "only one possible path into the future." That is what makes indeterminism important for free will.

## 2 Fischer's (Semi-)Compatibilism and Frankfurt-type Examples

With this in mind, let us turn to the other three essays. In chapter 2, John Fischer concedes that the *freedom to do otherwise* is *not* compatible with determinism, as libertarians like myself also believe. Fischer makes this concession in part because he is well aware of the intuitive appeal of the garden of forking paths image. (Indeed, Fischer is responsible for introducing Borges' image of forking paths into current free will debates.) Thus, he admits that "it is extremely natural and plausible – almost inevitable – to think of ourselves as (sometimes at least) having more than one branching path into the future." I agree. In addition, Fischer thinks the Consequence Argument makes a "highly plausible" (though not indisputable) case that, if causal determinism is true, agents could *not* have done otherwise than they actually did. Again, I agree.

But Fischer then makes a surprising move. For, while he concedes that the *freedom* to do otherwise is not compatible with determinism, he insists that *moral responsibility is* compatible with determinism because he believes moral responsibility does not require that agents could have done otherwise. (This is his *semi*compatibilism: *Freedom* requires the power to do otherwise, or *alternative possibilities* (the condition I called AP), but moral responsibility does not require alternative possibilities or AP; and hence moral responsibility is compatible with determinism, though freedom to do otherwise is not.) This is an unusual position. So we have to ask *why* Fischer thinks moral responsibility does not require alternative possibilities. For that is the key to it. The answer has to do in large part, though not entirely, with his appeal to "Frankfurt-type examples," which are discussed in his chapter and also in Pereboom's chapter. Those who defend Frankfurt-type examples, such as Fischer, believe these examples show that *moral responsibility does not require that agents could have done otherwise* and thus provide powerful reasons for thinking that moral responsibility is compatible with determinism. I think they are wrong about this. But let us look at their argument.

Frankfurt-type examples all have this structure: A controller or mechanism (Black, the neurosurgeon, in Fischer's example) is capable of preventing an agent (Jones) from doing otherwise. But the agent, Jones, goes ahead and does *on his own* what the controller wants (Jones votes Democratic). So the controller Black does not have to intervene to make Jones do what he wants and the controller does not intervene. In that case, say Fischer and other defenders of Frankfurt-type examples, we have every reason to believe Jones is responsible for voting as he did because he did it *on his own* without being coerced and Black played no role in it. But if Jones may be responsible for acting on his own in such circumstances, nonetheless he *could not have done otherwise* because the controller Black would not have let him. So, being *responsible*, they conclude, does not require that one *could have done otherwise* (AP) and therefore does not require the falsity of determinism.

One might even go on to imagine a "global Frankfurt controller" who is present throughout the entire lifetime of an agent, like Jones, but never intervenes in the agent's lifetime because the agent always does on his own what the controller wants. The agent might then be responsible for many acts done in his lifetime because the controller never intervened. Yet the agent could *never* have done otherwise in his entire lifetime because the controller would never have let him. So, the agent could be responsible even if he could never have done otherwise (and hence even if determinism was true).

What are we to say about these "Frankfurt-type examples"? If you are like me, your initial reaction is that they are pretty strange and esoteric examples on which to base a *general* belief that moral responsibility is compatible with determinism. Do they have anything to do with our ordinary world where

there probably are no Frankfurt controllers (at least not yet)? But let us put that worry aside, since philosophers are often into strange and esoteric examples. Of more importance, in my view, is the fact that these unusual Frankfurt-type examples don't really prove what they are suppose to prove anyway. In particular, they do not show that moral responsibility is compatible with *never* having the ability to do otherwise and so they do not show that moral responsibility is compatible with determinism. My reason for saying this has to do with an objection against Frankfurt-type examples that I put forward more than twenty years ago that was later developed independently by David Widerker and others. This objection to Frankfurt-examples is referred to as "the Kane/Widerker objection" by Pereboom in his essay. Here is my version of this objection:

> Incompatibilists and libertarians about free will like me believe that if we are to be ultimately responsible for being the way we are, then there must be some choices in our lifetimes that are undetermined right up to the moment when they occur. These undetermined choices are the "will-setting" or "self-forming" actions (SFAs) discussed in chapter 1 that are required at some points in our lives if we are to be ultimately responsible for forming our own wills. Now a Frankfurt controller faces a dilemma in trying to control these will-setting or self-forming choices. Since they are undetermined up to the moment they occur, a Frankfurt controller cannot be sure which way the agents are going to choose *before* they actually do choose. Thus, if the controller waits till the agents actually make one choice or the other, it will be *too late* to intervene and the agents may choose against the controller's wishes (since the will-setting choices are undetermined and may go either way). In that case, the agents may be *responsible* for acting on their own, but they will also have *alternative possibilities* at the moment of choice. By contrast, if the controller is to *ensure* that the agents will do what the controller wants, the controller must act *in advance* to *make* the agents choose as the controller wishes. In that case, the agents will indeed *not* have alternative possibilities, but neither will the agents be ultimately responsible for the outcomes. The controller will be responsible, since the controller will have intervened in advance to determine which outcome would occur.

In sum, the indeterminism required by will-setting or self-forming actions (SFAs) would "thwart" any potential Frankfurt controller. The only way such a controller could prevent alternative possibilities in the case of SFAs would be to intervene in advance to determine which choice would be made; and then the controller would be responsible for the outcome, not the agents.

But note that, as argued in chapter 1, we must make *some* SFAs or self-forming choices at some points in our lifetimes, if we are to be ultimately responsible for forming our own wills and hence for having free will in the

true libertarian sense. Otherwise, there would be nothing we could have *ever* done differently in our entire lifetimes to make ourselves different than we are. So if Frankfurt-type examples fail for SFAs, such examples do not show that moral responsibility is compatible with *never* having the ability to do otherwise in our entire lifetimes. Hence such examples do not show that moral responsibility is compatible with *all* of our actions being determined.

## 3 Pereboom's Frankfurt-type Example: Tax Evasion

In response, Fischer acknowledges that this sort of objection, which Pereboom calls the Kane/Widerker objection, poses a serious challenge to the claim that Frankfurt-type examples show moral responsibility is compatible with determinism. But Fischer adds that a number of new and more sophisticated Frankfurt-type examples have been devised by philosophers in recent years that have the promise of answering this objection and vindicating the compatibilists' claim that moral responsibility is compatible with determinism. As it happens, one of these new Frankfurt-type examples to which Fischer is referring was introduced by Pereboom and is presented in Pereboom's chapter 3. Pereboom is also a defender of Frankfurt-type examples. But he thinks the usual Frankfurt-type examples fall prey to the above "Kane/Widerker objection." Pereboom's new Frankfurt-type example is specifically designed to answer this objection. I do not think it does so. But it is instructive to see why.

According to Pereboom, the problem all these new Frankfurt-type examples must overcome if they are to answer the Kane/Widerker objection is this: A Frankfurt controller needs a *prior sign* that can *reliably* tell the controller *in advance* what the agent is going to do. Without such a prior sign, the controller is in a bind about whether or not to intervene, as we have seen. So Pereboom asks: How can there be a prior sign or "cue" that will reliably tell the controller in advance whether the agent is going to do what the controller wants, *even when the choice is undetermined*? Pereboom's subtle example, which he calls "Tax Evasion," is meant to show how this could be.

A fellow named Joe is inclined to evade taxes by claiming an illegal deduction. The only thing that will make Joe hesitate to choose to evade the taxes is if "he attains a certain level of attentiveness to moral reasons." Attaining this level of attentiveness will not *determine* that Joe will refrain from evading taxes, but it will allow Joe to have an (undetermined) libertarian free choice to either choose to evade the taxes (call that choice A) or to refrain for moral reasons (call that choice B). If Joe fails to attain this level of attentiveness, however, he will definitely choose to evade the taxes (A) because he will not have sufficiently attended to any other reasons (moral reasons) for doing

otherwise (B). Now enter a Frankfurt controller who is monitoring Joe's brain. If the controller sees in advance that Joe has failed to attain the required level of attentiveness, the controller will not intervene because he will know that Joe is going to make choice A on his own and Joe will not even consider choice B. But if the controller sees that Joe *has* attained the required level of attentiveness, the controller will intervene to *ensure* that Joe will make choice A and not B.

The difference in this example from other Frankfurt-examples, according to Pereboom, is that the occurrence of the "prior sign" or "cue," which tells the controller to intervene – i.e., Joe's attaining the required level of attentiveness – does not *determine* or guarantee in advance *which* choice Joe is going to make. For, if Joe attains the level of attentiveness, he will have an undetermined libertarian choice between A and B – perhaps even an SFA. But, alas, a familiar problem arises once we introduce the controller: For, the controller is not going to *let* Joe make that *undetermined* choice between A and B! If Joe does attain the level of attentiveness, the controller is going to intervene and prevent Joe from having that undetermined SFA between A and B *that Joe might otherwise have had*. So Joe will not get a chance to make a true SFA *either way* once the controller is in the picture. If Joe attains the level of attentiveness, the controller will intervene and make him choose A; and if Joe does *not* attain the level of attentiveness, he will choose A on his own because he will not have attended to any reasons (moral reasons) for doing B. Joe's choice in the second case will not be a "will-setting" SFA either because he will only have reasons to "set his will" on A and will not have attended to any good reasons to set his will on B.

Now Pereboom argues that in his example, unlike other Frankfurt-examples, it might be undetermined right up to the moment of choice (of A or B) whether Joe does or does not attain the required level of attentiveness to moral reasons. So it might be open until the moment of choice whether Joe will choose A or B. But I don't see how this solves the controller's problem. If Joe does *not* attain the level of attentiveness until the very moment in which he makes the choice of A or B itself, the controller will not be able to know whether or not to intervene until the choice is actually made; and then it will be too late. But if Joe does attain the level of attentiveness before the moment of choice, *even a very short time before*, so that the controller has just *enough time* to intervene and does intervene, then the controller will be responsible for the outcome and not Joe. So the old dilemma for all Frankfurt-type examples returns in full force: If the controller waits till the agent chooses A or B to find out what the agent is going to do, the agent will have alternative possibilities; and if the controller intervenes before the choice is made (even a short time before), the agent will not be responsible, the controller will be.

To sum up, my objection to Pereboom's example is the same as my objection to all Frankfurt-type examples: They do not work *for self-forming choices or SFAs* which must be will-setting and undetermined up to the moment they occur. If the controller is able to prevent the agent from doing otherwise, the choice available to the agent will not be an SFA. But we must make *some* SFAs or self-forming choices at some points in our lifetimes, if we are to be ultimately responsible for forming our own wills and hence for having free will in the libertarian sense I defend. Otherwise, there would be nothing we could have ever done differently in our lifetimes to make ourselves different than we are. Frankfurt-type examples therefore do not show that moral responsibility is compatible with *never* having the ability to do otherwise; and so they do not show that moral responsibility is compatible with determinism.

Returning for a moment to Fischer, recall that he concedes that *freedom to do otherwise* is not compatible with determinism; yet he nonetheless wants to hold on to the idea that *moral responsibility* is compatible with determinism. But if I am right in arguing that moral responsibility *sometimes* requires the *freedom to do otherwise* (in SFAs) at some points in our lives, then moral responsibility would also not be compatible with determinism (though it might be compatible with *some* morally responsible actions being determined.)

## 4 Indeterminism, Luck, and Chance

Those who believe that moral responsibility is compatible with determinism have one other move they could make in response to the preceding arguments. They may argue that self-forming choices or SFAs are simply impossible. If choices are undetermined right up to the moment they occur, as SFAs must be, one might argue, then their occurrence one way or the other would be merely a matter of "luck" or "chance" and the choices would not be free and responsible actions at all. This turns out to be the familiar "luck" or "chance" objection to libertarian accounts of free will, which was the "second prong" of the modern attack on libertarian free will mentioned in chapter 1: If free will is not compatible with determinism, it does not seem to be compatible with *indeterminism* either. Undetermined events happen by chance or spontaneously, so the argument goes, and are not controlled by anything, hence not controlled by the agents. We thus get the familiar charges that undetermined choices of the kinds that libertarians demand would be "capricious," "random," irrational," "uncontrolled," and therefore not really free and responsible actions at all.

I addressed many of these objections in chapter 1, where I argued that undetermined self-forming choices or SFAs of the kind required for free will

need not be "capricious," "random," "irrational," or "uncontrolled," and therefore can be free and responsible actions. But the "luck" and "chance" objections are stated in new forms by the other authors of this volume in their critiques of libertarian views. So I must address their further criticisms. In chapter 3, for example, Pereboom offers a lengthy and subtle defense of the "luck objection" against my view. "The luck objection," he says, claims that, for libertarian views like Kane's, "the relevant causal conditions antecedent to a decision . . . leave it open whether this decision will occur and the agent has no further causal role in determining whether [the decision] does [occur] . . . Accordingly [for libertarian views like Kane's] the agent lacks the control required for being morally responsible for the decision."

Now this statement is misleading in a number of ways. The first part of it is true. It is true that on my view (and most libertarian views) "conditions antecedent to a decision," such as the agents' reasons, motives, intentions, and other agent-involving states, "leave it open whether this decision will occur" or not. For, this is just another way of saying that the free decision or choice is not *determined* by the agent's antecedent mental and physical make-up, though it may be influenced by that make-up. But it is misleading to go on to say that, on my libertarian view, "the agent has *no further causal role to play* in determining whether" the decision occurs. This claim is misleading in several possible ways, which I will address in turn. First, the claim is misleading to the extent that it suggests that, on my view, the agents are not the *causes* at all of their undetermined self-forming decisions or SFAs; that all the agents do is somehow "let them happen." For that is not true. It overlooks the role that "efforts of will" play in my theory.

Agents on my view *cause* or *bring about* their self-forming choices or SFAs and they do so voluntarily by making efforts of will. These efforts of will are in turn causally influenced by the agent's motives, reasons, and other states of mind, but their outcomes are not determined by these motives and reasons. The efforts may fail. But if the efforts *succeed*, as I argued, the *agents* will have voluntarily *brought about* the choices *as a result of their efforts* and hence can be held responsible for doing so. Like the husband who swings his arm down on the table in an effort to break it, his motives or reasons influence his effort, but do not determine it. He may fail to break the table. But if his effort succeeds, *he* (the agent) will have brought about the breaking of the table and he will be responsible for it.

That's why the husband's excuse to his wife – "chance broke the table, not me" – was so lame. While chance was causally involved, chance was not the cause of the table's breaking; *he* (the agent) was, by making an *effort* to break it. The chance played merely an interfering role, like the noise in the electrical lines that were transmitting the Morse code message. If the message gets through, despite the electrical noise, the cause of its getting through is not

the interfering noise, but the effort of the sender of the message to get the message through despite the interference. And so it is, I suggested, with the efforts leading to self-forming choices. These efforts of will are *mental* efforts, of course, that are realized in the higher cognitive processing of the brain rather than in overt actions such as the swinging of an arm. But the SFAs that result from these mental efforts are also the *achievements* of goal-directed activities that might have failed due to chance, but did not. When the agents succeed in these mental efforts, they are responsible for the outcomes (in this case, the *choices*), just as the husband was responsible for the outcome of his effort when he succeeded.

One can see from these remarks why it is misleading to say that on my view agents have "no causal role to play" in determining whether an SFA occurs or to say that agents could not be morally responsible when the choices result from their mental efforts even if the choices are undetermined. But, to be fair to Pereboom, he is actually saying something stronger than this. He is saying that, to be responsible, agents must play some "*further* causal role" in determining whether self-forming choices occur *over and above* the role played by the agents' "reasons, motives, intentions, and other agent-involving states." But this further claim is also mistaken, I have argued, when efforts of will are brought into the picture. For in *making efforts* of will to choose in terms of their reasons and motives, agents *do* play a causal role in bringing about their choices over and above the causal role played by their reasons, motives, intentions, and other mental *states* alone. In other words, it's not as if the agents sit back and watch while the reasons or motives cause the choice. It is rather that, by making efforts, the agents actively bring about the choice *for* the reasons and motives. Efforts are different from *desires* and other motives in this respect, contrary to what Pereboom suggests, because efforts are actions of the agents and not merely *states*. Agents can be held responsible for what they bring about by their efforts, even if the efforts might have failed due to indeterminism, just as the husband and the Morse code sender can be responsible when they succeed in their efforts.

But don't we also have to be responsible for the efforts we make, if we are to be responsible for the choices they produce? Yes, and in the vast majority of SFAs humans normally make, including the businesswoman's, responsibility for the efforts they make comes from two sources. First, responsibility comes from the character and motives influencing the efforts, which in most normal adult humans have been built up by many past self-forming choices in their lifetimes; and, second, an additional responsibility is added by the present effort itself by virtue of the agent's endorsing its outcome when it succeeds. Thus, responsibility accumulates in human beings as they get older and build up a backlog of self-formed character. The only exceptions to this twofold source of responsibility are the earliest SFAs of childhood in which

a backlog of *self-formed* character does not exist. In these earliest SFAs, all the responsibility is thus in the effort itself and the endorsement of its outcome by the agent when it succeeds. But precisely for this reason the responsibility for these earliest SFAs of childhood is not as great as later ones; in fact, in the earliest SFAs responsibility is minimal. That is why we hold very young children to be far less responsible than older ones and adults.

In fact, I believe the earliest SFAs of childhood have a probative (or probing or learning) character to them. Young children are often "testing" what they can get away with (the limits) and what consequences their behavior will have on them and others. (That is one reason why childrearing is so exhausting.) A toddler may think: Should I take a cookie from the jar or should I obey Mommy and wait for dinner? When he has to stop and think about it for the first time rather than just doing what he wants, he is facing his first SFA. He is torn between pleasing his mother and having the cookies. His mother can and should hold him responsible, if he takes the cookies. But the responsibility is minimal, if it is his first SFA, because the child is only beginning to form his character. So perhaps he will be sent to his room or not get cookies for dinner. A more severe punishment would be inappropriate in the beginning. But the mother will hold him responsible in some small way, if she is a wise parent, for she knows that this is how his character will be formed. Thus is character slowly built up by how children respond to these earliest probing SFAs of childhood. If a three year old is told not to take more than his share of cookies, but tries to do so anyway the next time, the child is responsible, but not as responsible as when he does it a second, third, and fourth time and it becomes a pattern of behavior.

## 5 The "No-Further-Power" Objection

Pereboom concedes some of the points I have just made about undetermined decisions and SFAs, but he does not think they quite settle the matter; and so he introduces one further objection. Thus, he says:

> Kane . . . argues that decisions can be undetermined and yet have many features indicative of agent control and moral responsibility. Undetermined decisions could still be made . . . for reasons; agents might make them for reasons, rather than by mistake, accident, or chance; and agents may want to make them for these reasons rather than any others. Agents might not be coerced or compelled in making undetermined decisions, and in making them they might not be controlled by other agents or circumstances . . . The fact that indeterministic action can have these characteristics does show that it allows for significant control in action. (p. 105)

But Pereboom wonders whether these conditions are *sufficient* for *libertarian* moral responsibility. For he notes that *compatibilists* appeal to similar characteristics to defend their view. Compatibilists say, for example, that even if decisions were *determined*, the decisions "could still be made for reasons, agents might make them for reasons rather than by mistake, accident or chance . . . Agents might not be coerced or compelled in making" them, and so on. Pereboom thus argues that, while my libertarian view may "allow *as much* responsibility-relevant control as compatibilism does," it does not give us any *further* power or control over our decisions or choices than compatibilists give us. But some further power or control over our decisions than compatibilists give us, it would seem, is required to account for true *libertarian* responsibility.

I call this the "no-further-power-or-control-than-compatibilists-give-us" objection to libertarian accounts of free will like mine, or the "no-further-power" objection, for short. (Pereboom notes that this objection has also been made by others, such as Randolph Clarke.) The objection is important because it is surely true that we must have more power or control over our choices than compatibilists could give us in a thoroughly *determined* world, if we are to have true *libertarian* free will and responsibility. But I think this "no-further-power" objection also fails because there is very clear sense in which the view I have described *does* give us more power than compatibilists can give us in a determined world. To see why, we need to review some key points from chapter 1.

I suggested that will-setting or self-forming choices (SFAs) occur at those difficult times of life when we are torn between competing visions of what we should do or become. In all such cases, we are faced with competing motivations and have to make efforts to overcome the temptations to do something else we also strongly want. When we do decide under such conditions, the outcome can be willed either way we choose, rationally and voluntarily, owing to the fact that in such self-formation, the agents' prior wills are divided by conflicting motives. As a consequence, in such circumstances, the agents have what I called *plural voluntary power or control* over their options in the following sense: They able to bring about *whichever* of the options they will, *when* they will to do so, *for* the reasons they will to do so, *on* purpose, rather than accidentally or by mistake, *without* being coerced or compelled in doing so or willing to do so, or being otherwise controlled in doing or willing to do so by other agents or mechanisms.

Each of these conditions can be satisfied by SFAs, like the businesswoman's, as I argued in chapter 1. The reason is that, at any given time, either of the efforts she is making (to stop and aid the victim or to go on to her meeting) might succeed. She would then bring about that choice as a result of her effort; and she might succeed at either one at any time during

her deliberation. The conditions for plural voluntary control just stated can be summed up by saying that the agents are able to choose *in more than one way* voluntarily, intentionally, and rationally, no matter which alternative is chosen. This is what is meant by saying the choices involved are "will-setting" and satisfy the "plurality conditions": The agents can set their wills one way *or* another in the *act* of deciding.

Now return to the "no-further-power" objection. The objection is that my libertarian view does not give us any further power or control over decisions than compatibilists give us, which is not enough for true *libertarian* responsibility. But I submit that, if agents can exercise plural voluntary power or control in the above sense over more than one undetermined alternative, as I have argued they can in SFAs, then the agents *do* have more power than compatibilists can give us in a determined world. For the most compatibilists can say of agents *in a determined world* who choose voluntarily, intentionally, and rationally is that the agents may have chosen *otherwise* (or in some other way) voluntarily, intentionally, and rationally, only *if* the past or the laws of nature had been different in some way (if the agents had different thoughts, desires, beliefs, etc.). Compatibilists cannot say that agents may have chosen otherwise voluntarily, intentionally, and rationally, given the *actual* laws of nature and the past *as it actually was* at the moment of choice.

I submit that such a *plural* voluntary power that may be exercised in acting *or* acting otherwise, given the laws and the past of the *actual* world at the moment of action, *is* further power than the merely *hypothetical* power to do otherwise that compatibilists can give us in a determined world. Compatibilist power to have done otherwise may have been exercised only if the past or the laws *had been different in some way*; and agents do not have the further power at the moment of action to change the actual past or the laws of nature by their present actions. (If we could now actually *change* the past or make false the laws of nature, things would be different. But those who make the "no-further-power" objection, such as Pereboom and Clarke, are not saying we do have such a power.)

Furthermore, not only is such a plural voluntary power a further power that compatibilists cannot give us in a determined world, but I submit that it is just the *kind* of power that libertarians have always demanded for free will and moral responsibility – that is, a power that can be voluntarily (non-coercively), intentionally (purposefully), and rationally (for reasons) exercised *in more than one way* here and now, given the *actual* laws of nature and the *actual* past at the moment of action.

So I think this "no-further-power" objection ultimately fails. Plural voluntary power, given the actual past and laws of nature, is more power than compatibilists can give us in a determined world. But the "no-further-power" objection does have an initial appeal and it is instructive to ask why it has

this appeal. The reason, I believe, is that it is hard to see how adding *indeterminism* to the picture of free actions, as libertarians must do, could give persons *more* power or control rather than less. Indeterminism, it seems, would diminish our control over what we are doing rather than giving us more control, since it is a hindrance or an obstacle to our realizing our purposes.

This is an important point, but I also addressed it in chapter 1. In the case of the businesswoman (and SFAs generally), I argued, indeterminism *is* functioning as a hindrance or obstacle to her realizing one of her purposes – a hindrance or obstacle in the form of resistance within her will which has to be overcome by effort. If there were no such hindrance – if there were no resistance in her will – she would indeed in a sense have "complete control" over *one* of her options. There would no competing motives standing in the way of her choosing it and, therefore, no interfering indeterminism. But then also, she would not be free *to rationally and voluntarily choose the other purpose* because she would have no good competing reasons to do so. Thus, by being a hindrance to the realization of *some* of our purposes, indeterminism paradoxically opens up the genuine possibility of pursuing other purposes – of choosing or doing *otherwise* in accordance with, rather than against, our wills (voluntarily) and reasons (rationally).

## 6 Vargas and Revisionism

Manuel Vargas's critique of libertarian views of free will in chapter 4 takes a different line in some respects than the critiques of Fischer and Pereboom. First, Vargas says many things that libertarians like me can readily agree with. He argues, for example, that "common sense is incompatibilist, that is, . . . common sense has elements that require a picture of agency whose commitments could not be satisfied in a determined world." Vargas also believes the Consequence Argument makes a powerful case for the incompatibility of free will and determinism. He is also suspicious, as I am, of compatibilist arguments, such as those involving Frankfurt-type examples, that claim to show that moral responsibility is compatible with determinism. Finally, Vargas does not believe that libertarian accounts of free will such as mine are incoherent or self-contradictory or unintelligible, as many critics of libertarianism believe.

Why then does Vargas think we should give up libertarianism and "revise" our views of free will and responsibility in a compatibilist direction? Briefly put, he doubts the "empirical plausibility" of libertarian accounts of free will. He thinks the scientific evidence can and will likely show them to be false. One reason he thinks this is that libertarian theories of free will are

empirically "more demanding" theories than the alternative views of free will (compatibilism, semicompatibilism, hard incompatibilism, revisionism). What does Vargas mean by this? Well, roughly, libertarians demand that there be some indeterminism in nature, so that physical determinism must be false. Further, the indeterminism that libertarians require cannot be just anywhere in nature; some of it must be in the brain, where it can have some effect on human decision-making. Further, as Vargas suggests, all libertarian theories must demand this, not just my theory. If the brain were a completely deterministic mechanism there would be no room in nature for libertarian free will to fit in, even if one were a dualist about mind and body, like Descartes.

But now note that these empirical requirements are *not* made by other alternative theories on free will. Compatibilists of the standard variety believe that free will and moral responsibility are both compatible with determinism. So, on their compatibilist view, both free will and moral responsibility could exist in a thoroughly determined world and also if the brain were a deterministic mechanism. Semicompatibilists like Fischer believe this also about moral responsibility (though not about the freedom to do otherwise). This is why Fischer argues that it is an advantage of compatibilist views like his that they would not be threatened by future developments of science. In a striking image, he says that if a headline saying "Scientists have discovered that determinism is true" were to appear in a distant future newspaper and turn out to be true, his view and other compatibilist views would not be threatened. By contrast, libertarian views of free will would turn out to be false and libertarians would have to abandon their view.

Note further that Pereboom's "hard incompatibilism" would obviously also not be threatened by Fischer's future headline because Pereboom thinks that libertarian free will and libertarian moral responsibility do not exist anyway. Libertarian views, he argues, are either impossible or empirically false. If we believed we do not have libertarian free will anyway, as Pereboom and other hard incompatibilists believe, Fischer's future headline would also not trouble us. Finally, Vargas's "revisionist" view would also not be threatened by Fischer's future headline. Vargas agrees with Pereboom that we cannot have libertarian free will. But rather than accepting what he takes to be Pereboom's pessimistic conclusion that we must "live without free will," Vargas suggests that we should revise our commonsense views in a compatibilist direction. We should hang on to notions of freedom to do otherwise and moral responsibility, but revise our commonsense thinking about them so that they turn out to be compatible with determinism.

In summary, when Vargas says that libertarianism is a "more demanding" theory empirically than the other alternative theories of free will, this is what he seems to mean: Future science could show libertarianism to be false in

ways that would leave the other views still standing. On this point, I think he is right. But, as a libertarian I would respond as follows: Yes, the libertarian view is demanding: It could turn out to be false. Future scientific research into the cosmos and human nature could show it to be false (or true). There are no a priori guarantees or proofs independent of experience that we have libertarian free will.

But I would argue that if the alternative views on free will discussed in this volume are immune to such scientific refutation, it is because what these alternative views give us are merely "watered down" notions of free will and responsibility (or, in the case of hard incompatibilism, no free will or true moral responsibility at all). *If you want something as important as libertarian free will and moral responsibility, then you are going to have to take your chances with the scientific evidence.* And if you don't want to take those chances, you will have to accept some watered-down versions of free will and moral responsibility (or none at all).

For example, the "freedom to do otherwise" that standard compatibilists offer us (we would have done otherwise, *if* we had wanted or chosen otherwise, that is, if the past had been different in some way) is indeed compatible with determinism. But such a freedom to do otherwise seems to me, as it did to Kant, "a wretched subterfuge" for the real freedom to do otherwise represented by the garden of forking paths. And notice that Vargas also needs some "compatibilist" notion of the freedom to do otherwise in his revisionist theory. He says the "can" in "can do otherwise" must be interpreted so that it is compatible with determinism. Vargas acknowledges that the standard compatibilist analysis of "could have done otherwise" ("we would have done otherwise, *if* we had wanted otherwise") is subject to serious objections. He thinks some better compatibilist analysis can be found, but he does not offer a developed alternative. That is future work for him to do in developing his theory; and it will be hard work, in my view. For every alternative compatibilist interpretation of the ability to do otherwise I have yet seen is a wretched subterfuge for the real freedom to do otherwise represented by the garden of forking paths.

Turning to Fischer's semicompatibilism, it would also not be refuted if science were to prove determinism was true or that the brain is a deterministic mechanism. But that is because, on Fischer's view, moral responsibility does not require the freedom to do otherwise (in the sense of forking paths) *at all*. For Fischer, "guidance control" is enough for moral responsibility and guidance control does not require that we could have done otherwise than we actually did. If science cannot refute such a view, that is because what such a view offers is also, as I see it, a weak and watered-down notion of moral responsibility. Finally, Pereboom's view would also obviously not be threatened if scientists proved determinism was true or that the brain was a

deterministic mechanism because he denies we *have* free will or moral responsibility in the true senses anyway, which he thinks are incompatible with determinism.

So, yes, I concede to Vargas that libertarianism is a more empirically demanding theory in this sense than these others. Future scientific research into the cosmos and the brain could show it to be false. But that doesn't mean that, as a libertarian, I am prepared to jump over to one of these alternative views, which offer what I regard as pallid substitutes (if not subterfuges) for free will, or no free will at all. I prefer to wait, thank you, till science does prove determinism is true or that the brain operates on strictly deterministic laws; or at least till a good deal more evidence on these matters is in. If I do ever read Fischer's future headline and it is true, I would give up my libertarian view and perhaps go over to one of these other views. I think empirical evidence matters. But I don't know which of these other views I'd go to. For someone with libertarian intuitions like me, it would be like being asked whether I wanted to live in the desert, in the middle of the jungle or at the South Pole. Well, I don't like any of the options. Do I have to choose *now*? Can I spend a few weeks in Hawaii while I think about it?

Let me say a final word about the empirical evidence. Some persons, like Vargas and Pereboom, are inclined to think the empirical evidence is already decisive against libertarianism. But I think it is way too soon to prejudge these empirical matters. One way science might show libertarian views to be false is by showing that determinism is true. But it is far from clear today that universal physical determinism is true. Most scientists and philosophers believed it to be so in the nineteenth century, but then quantum physics came along, introducing elements of indeterminism or chance into scientific accounts of the physical world. There are still debates about whether quantum physics really is indeterministic or whether it might be superseded by some future theories of the physical world that are deterministic. But, in a recent survey of the latest theories of physics, including string theory, the eminent philosopher of science John Earman concludes that we are a very long way from being able to decide definitely whether determinism is or is not true of the physical world – and we may never know for sure.

The situation is more complicated in the case of the brain sciences. But even there, I prefer to keep an open mind on where things might go. The cognitive and neurosciences have made tremendous strides in the past few decades. (I make use of some of their new ideas, such as parallel processing, in my account of free will.) But I don't think we can yet say with full confidence where they will go in future on these matters. Vargas is correct in saying that "there are at present no accepted scientific models of indeterministic brain events" and also in saying that at least some existing theories about "indeterminism-amplifying aspects of the brain (of Penrose and Hameroff,

for example) have been widely rejected by neuroscientists, philosophers and mathematicians." But it is misleading to add that "virtually all brain science proceeds on the basis of thoroughly deterministic explanations." Vargas seems to have in mind that many brain scientists proceed on the assumption that deterministic explanations will eventually be found for all functioning of the brain. But this is an assumption that says as much about their scientific predilections as about the finality of the present evidence.

As Mark Balaguer (unpublished paper) has recently argued, the claim that current "neuroscience treats all neural processes as deterministic is straightforwardly false. Current neuroscientific theory treats a number of different neural processes probabilistically, and any decent textbook on neuroscience will point this out. For instance, synaptic transmission [the synapses are those places where signals are transmitted from one neuron to another] and spike firing [the firings of individual neurons] are both treated probabilistically. One textbook (Dayan and Abbott, 2001) puts these points as follows:

> [Synaptic] transmitter release is a stochastic process. Release of transmitter at a presynaptic terminal does not necessarily occur every time an action potential arrives and, conversely, spontaneous release can occur even in the absence of [the arrival of] . . . an action potential. (p. 179) . . . Because the sequence of action potentials generated by a given stimulus varies from trial to trial, neuronal responses are typically treated statistically or probabilistically. For example, they may be characterized by firing rates, rather than as specific spike sequences. (p. 9)

"It is worth noting," Balaguer adds, "that some aspects of the indeterminacies of both of these processes are caused by the indeterminacy inherent in another process, namely, the opening and closing of ion channels, which are essentially little gates that let charged ions in and out of cells. Now, to be sure, by treating these processes probabilistically, neuroscientists do not commit themselves to the thesis that, in the end, they are genuinely indeterministic. But the important point here is that they aren't committed to determinism either. The question of whether these processes are genuinely indeterministic simply isn't answered by neuroscientific theory."

Balaguer goes on to quote several neuroscientists, including Dayan, one of the authors of the text just cited, who says that "people would argue that there are good thermal reasons to think that [the opening and closing of ion channels] is truly random. Thus, short of philosophical debates about hidden variables for all forms of randomness in physics, this is some fundamental randomness for which people nearly have evidence." And Sebastian Seung, a neuroscientist at MIT, says that "The question of whether [synaptic transmission and spike firing] are 'truly random' processes in the brain isn't really a

neuroscience question. It's more of a physics question, having to do with statistical mechanics and quantum mechanics" (quoted from correspondence by Balaguer, ibid., p. 34).

None of this proves that indeterminism plays a significant role in the brain. We are a long way from showing that. But neither does it show that the brain operates on only or strictly deterministic laws. I prefer to keep an open mind on such issues and am not about to go over to some other view until far more evidence is in. As it happens, on my libertarian account of free will, one does not need large-scale indeterminism in the brain, in the form, say, of *macro*-level wave function collapses (in the manner of the Penrose/Hameroff view mentioned by Vargas). Minute indeterminacies in the timings of firings of individual neurons would suffice, because the indeterminism in my view plays only an interfering role, in the form of background noise. Indeterminism does not have to "do the deed" on its own, so to speak. One does not need a downpour of indeterminism in the brain, or a thunderclap, to get free will. Just a sprinkle will do.

# 6

# *Response to Kane, Pereboom, and Vargas*

*John Martin Fischer*

## 1 Response to Kane

Kane helpfully distinguishes two requirements for free will: the alternative possibilities requirement (AP) and the condition of "ultimate responsibility" (UR). He believes that (UR) is the fundamental worry for compatibilists, and that it is even more important than (AP). Further, Kane contends that, although UR "does not require that we could have done otherwise for *every* act done of our own free wills, . . . it does require that we could have done otherwise with respect to *some* acts in our past life histories by which we formed our present characters" (Kane, p. 14). Kane applies this idea to cases such as Martin Luther, who alleged that he literally couldn't have done otherwise (than break with the Church in Rome); Kane says that we can (and, indeed, must) trace Luther's moral responsibility to past acts to which there were genuine alternative possibilities open to Luther. Kane contends that if there were never alternative possibilities open to us (at *any* points in our lives), then "*there [would be] nothing we could have ever done to make ourselves different than we are*" and thus we would not satisfy (UR).

I agree with Kane that we can distinguish between an alternative possibilities condition on moral responsibility and a sourcehood condition. But I do not see why a suitable sourcehood condition should entail any sort of alternative possibilities condition (even a "tracing" condition – a condition that reaches back into the past). Kane seeks to argue that (UR) entails a version of (AP). The argument is that (UR) entails the existence of "will-setting" actions at some points in our lives, which in turn entails that some of our actions satisfy the "plurality conditions." But the plurality conditions entail alternative possibilities at some points in our lives. It is to Kane's credit that he seeks to

*argue* for the putative connection between (UR) and some version of (AP). But I am unconvinced. I do not see why "will-setting" should be any different from ordinary action with respect to the requirement of alternative possibilities. For example, why can't one have Frankfurt-examples at the stage of "will-setting"; why insist on the plurality conditions at the level of the will any more than the level of "external action"? Why the asymmetry?

I believe that it is better to insist on a sharp and pure distinction between the source condition and the alternative possibilities condition. When one acts freely, one must be the source of one's action in some suitable sense. But, of course, it is contentious whether sourcehood is consistent with causal determination. On my view, if Luther does not freely break with the Roman Church (and it is unclear whether this is literally true), his moral responsibility for breaking from the Roman Church can indeed be traced to previous free actions: it can be traced to past instances of choosing and acting freely. But choosing and acting freely do not require freedom to choose or act differently; guidance control does not entail regulative control. One can explain the "Luther" examples via this sort of tracing approach (in which one reaches into the past to find instances of acting freely or guidance control), just as one can explain moral responsibility for the consequences of drunk driving or behavior under the influence of drugs. There is no need to abandon the guidance-control model; there is no need to posit alternative possibilities at any point in the sequence issuing in behavior (or the lives of morally responsible agents).

## 2 Response to Pereboom

A major challenge for a proponent of the compatibility of causal determinism and moral responsibility is to characterize the distinction between "ordinary" causal determination and "special," responsibility-undermining causal determination. Pereboom's four-case argument involving Professor Plum nicely brings out this challenge. I have contended that Plum meets the minimal conditions for guidance control and moral responsibility in all four cases; thus, I do not seek to distinguish the cases with respect to moral responsibility. But clearly Plum is not blameworthy in cases one and two, given the provenance of the motivational states that lead to his behavior. On my approach, moral responsibility is the "gateway" to such notions as praiseworthiness and blameworthiness; an individual is morally responsible to the extent that he is an *apt candidate* for the reactive attitudes and moral praise and blame, but further conditions have to be met to secure justification for the *application* of such attitudes (and for moral praise and blame). An individual who exhibits guidance control has satisfied the freedom-relevant

condition linked to moral responsibility; he is thus importantly different from a nonhuman animal or a mere automaton, or a severely impaired human being. But it does not follow that it would be justified to morally praise or blame the individual or to apply any of the reactive attitudes.

It is relatively uncontentious that someone could perform a morally neutral action, be morally responsible, but be neither morally praiseworthy nor blameworthy. But Pereboom does not see how an individual could do something wrong, be morally responsible (in the sense of moral responsibility relevant to the traditional debates about the relationship between causal determinism and moral responsibility), but not blameworthy. I agree that it is not straightforward to produce cases of this sort. But consider a scenario in which there has been substantial and recurrent physical and emotional abuse by a husband of his wife over many years. The wife has tried to leave this toxic and abusive relationship, but she has not been able to summon the strength. Finally, after her husband has begun (yet again) to beat her cruelly and brutally, she shoots him to death. Whereas these cases are complex and controversial, I would suggest that it is plausible that the wife has done the wrong thing in killing her husband, that she is morally responsible (given that the story is filled in compatibly with her exhibiting guidance control), and yet she is not morally blameworthy for what she has done. As with the case of Professor Plum, the causal history of the motivational states issuing in the behavior is relevant to ascriptions of blameworthiness; whereas both moral responsibility and blameworthiness require certain sorts of histories, the precise historical conditions are different.

Consider, similarly, a "drug-runner" of the sort depicted in the film, *Maria Full of Grace.* Suppose, more explicitly, that the individual was born to poverty, and was under considerable pressure to transport illegal drugs to America. Again, we can stipulate that the pressures were considerable, but that they fell short of issuing in compulsion. Here, as with the example of the abused wife, I am inclined to say that the individual is morally responsible, has done the wrong thing, but is not blameworthy.

On my view, judgments of moral responsibility and (say) blameworthiness are two separate, but obviously related moments in our evaluation of behavior. To establish moral responsibility is to show *eligibility* for certain responses; but such eligibility does not in itself imply that the responses ought to be made in a particular context. Consider, as an analogy, that a student becomes "UC-eligible" – eligible to be admitted to a University of California campus – in virtue of meeting certain academic standards. But it does not follow merely from a student's meeting these basic standards that a particular University of California campus must accept the student, or that the campus ought, all-things-considered, to accept the student. Each campus selects from the eligible students – but each campus has further criteria it may apply in addition to the threshold or minimal conditions involved in UC-eligibility.

Similarly, consider a context of legal responsibility. (Of course, the context is very different from that of university admissions, and I do not wish to suggest any parallels other than the structural point about the two separate, but related moments in the pertinent evaluation.) Consider the recent trial of Zacarias Moussaui. Moussaui was first convicted of aiding the 9/11 conspirators, and the penalty phase of the trial was particularly relevant. This phase had two separate (but related) moments. In the first, Moussaui was deemed *eligible* for the death penalty. In the second phase, the jury sought to determine whether, given that he is eligible, it is justifiable, all-things-considered, to apply the death penalty. They decided not to recommend the death penalty. The structure is similar to that of evaluations of moral responsibility and blameworthiness, on my approach. No matter how boring one's lecture classes are (of course, I'm not talking about philosophy classes), I doubt that anyone would deem matriculation as even remotely parallel to the death penalty! But my point is simply that we have structurally parallel phenomena in the legal context and the context of university admissions – an eligibility-judgment and a subsequent judgment (requiring the satisfaction of additional criteria). It is not implausible that judgments pertaining to moral responsibility and blameworthiness have a similar structure.

Michael McKenna has pointed out that we shouldn't focus solely on the "hidden causes" posited by causal determination; after all, we can equally focus on "agential properties," and when we do so, we are less inclined to be concerned about (say) Professor Plum's moral responsibility. Pereboom says:

> At the same time, he [McKenna] allows that drawing greater attention to the hidden causes and their deterministic nature could generate the intuition that Plum is not morally responsible. But given that each of these two strategies is equally legitimate, the result will be a stalemate. In response, I advocate drawing *equal* attention to the sorts of agential properties that typically serve as a basis for ascribing responsibility, and to the hidden causes and their deterministic nature by means of the four cases, and then let the intuitions fall where they may. (Pereboom, p. 101)

But I believe that we need to distinguish two sorts of "hidden causes." (The use of the term, "hidden cause," may have an unfortunate rhetorical effect of stacking the deck here.) One sort of hidden cause is certainly problematic – it would be at war with the relevant "overt" properties. That is, it would disable or bypass or hinder the operation of the overt properties. An example would be a subtle defect in an engine that impairs the functioning of the carburetor (right – we are dealing with an older car here – one with a carburetor); perhaps this "hidden" defect even disables the carburetor entirely. Another sort of "hidden cause" is simply the set of constituents of the overt properties – these

are the more specific or concrete ways in which the overt properties are *instantiated*. Such a cause is not "at war" with the overt properties – quite the contrary. Here we would be thinking of the specific materials of which the carburetor is made, which allow it to function properly.

Now surely if we posit the first sort of "hidden cause" of the relevant agential properties, this will threaten our view of ourselves as unimpaired, properly functioning agents. But nothing in the doctrine of causal determinism implies that there are hidden causes of the first sort; indeed, causal determination is entirely consistent with the presence solely of the second sort of hidden cause, and there is nothing obviously problematic about the second sort of cause. Perhaps, more carefully, there is nothing similarly problematic about the second sort of hidden cause; when I focus on such a cause, I find no inclination to conclude that there cannot be the unimpaired and proper functioning of the agential properties that underwrite guidance control and moral responsibility. Indeed, such "hidden causes" make the relevant agential properties work!

## 3 Respone to Vargas

In the context of expressing (admittedly legitimate) concerns about the employment of Frankfurt examples, Vargas says, "I remain skeptical (perhaps more than most) that we will be able to show that alternative possibilities aren't a deep and pervasive part of thinking about freedom and responsibility" (Vargas, p. 136). I am inclined to agree that alternative possibilities are indeed a "deep and pervasive" part of both pre-reflective and more reflective thinking about freedom and responsibility. But I deny that they are an *ineradicable* part of our theorizing – our reflective analysis of such topics. Because the Principle of Alternative Possibilities expresses such a plausible and attractive idea, and because we typically think of ourselves as selecting a path into the future (where there is more than one such path available), semicompatibilism is a significant revision of ordinary, commonsense thinking – as well as standard philosophical reflection on freedom and moral responsibility. I do not deny that alternative possibilities are a presupposition of commonsense as well as philosophical analysis; rather, I seek to explain how we can offer a subtler, more refined analysis which dispenses with the requirement of metaphysical access to alternative possibilities, but also preserves and explains the connection between freedom (of an appropriate sort – guidance control) and moral responsibility.

To be a bit more explicit. I distinguish between the concept of moral responsibility and the conditions of its application. With regard to the concept, I do not consider myself a revisionist; I attempt to understand a robust,

ordinary notion of moral responsibility. But my account of the conditions of its application are significantly revisionary (perhaps even revolutionary), insofar as I think it can apply even in contexts in which an agent has never had genuine access to metaphysical alternative possibilities. Similarly, deliberation and practical reasoning – understood as we ordinarily understand these notions – can take place even in the absence of access to really available alternatives, despite what is certainly the "deep and pervasive" commonsense view.

Vargas wonders about McKenna's "normative relevance" approach to analyzing the Frankfurt-cases. He also suggests that the "complexity" of some of the Frankfurt-cases should make us worried about whether they are really pointing us to something clear and important. I think these are reasonable worries, but allow me to offer a few reflections. I too lament the fact that some of the literature surrounding the Frankfurt-cases involves argumentation that is perhaps excessively intricate; indeed, Harry Frankfurt recently told me that this literature is a "young person's sport." But I believe that we can get a kernel of truth from the examples without having to delve into all of the complexities; I think there is a moral to the stories.

It seems to me that the Frankfurt-cases point us to something simple and important: moral responsibility depends on how the actual sequence issuing in the relevant behavior unfolds, and *not* on whether the agent has metaphysical access to alternative possibilities. The examples should elicit a strong (but defeasible) intuitive judgment that the lack of access to alternative possibilities is *irrelevant* to moral responsibility. The famous economist, Kenneth Arrow, employed a constraint on "social welfare functions" which he called "the independence of irrelevant alternatives"; it was designed to rule out a social decision based on certain intuitively *irrelevant* alternatives. Similarly, Frankfurt's cases can be seen as designed to show that the lack of access to alternative possibilities is *irrelevant* to moral responsibility. When I consider relatively simple versions of the cases, I form a strong and clear judgment to this effect.

Now, of course, this sort of judgment is defeasible, and it is obviously difficult to defend it in light of various objections. Much of the intricate literature surrounding Frankfurt-examples seeks to defend the initial intuitive judgment against objections, and it is often thought that a proper defense requires the presentation of a Frankfurt-case in which it is indisputably and uncontroversially true that the agent is morally responsible and lacks *any* alternative possibilities. Perhaps it is correct that an ideal and completely convincing defense would provide just such a case, but I believe it is a mistake to suppose that the cases cannot play a useful and instructive theoretical role, even in the absence of this sort of defense. I believe that the role of the examples is to render a certain view plausible and attractive, where this view can then be seen to be appealing – and defended – on independent grounds. As I have always emphasized, the use of thought-experiments, such as the

Frankfurt-examples, should be one part of an overall argument for semicompatibilism that also employs other ingredients. Strawson-style arguments, Dennett-style arguments, and Frankfurt-style arguments can be seen to triangulate on the same conclusion: semicompatibilism.

I believe that the normative relevance point is a special case of the idea that metaphysical access to alternative possibilities is *irrelevant* (at the most penetrating level of analysis) to moral responsibility. After all, mere flickers of freedom – alternative possibilities without any voluntariness – do not seem to be relevant; in a context in which there are by hypothesis no alternative possibilities and no moral responsibility, one makes *no progress* toward getting moral responsibility simply by adding mere flickers of freedom. (To suppose otherwise would be to believe in a kind of alchemy.) Similarly, how can adding a normatively irrelevant possibility – a "strange or trivial possibility" – such as the freedom to roast a chicken, help to get from no responsibility to the existence of moral responsibility? Again, the contention that adding *only* the genuine availability of roasting a chicken is relevant – that it helps to generate moral responsibility from no responsibility – would seem to involve a kind of alchemy. The basic insight one gets from consideration of Frankfurt-cases – the moral of the stories – is that what matters to moral responsibility is the actual sequence, and not access to alternative paths. Different ways of defending this insight can be seen as different specific ways of bringing out the fundamental point that alternative possibilities are *irrelevant* to ascriptions of moral responsibility. Finally, if one is tempted to seek decisively and comprehensively to refute objections, there can emerge an alienating and unattractive complexity; but perhaps one should not succumb to the temptation. The Frankfurt-examples should be used in moderation, and should be part of an overall strategy for healthy philosophical analysis of freedom and responsibility!

Vargas points out that there is "another route to incompatibilism" – the source route (rather than the alternative possibilities route). From the very beginning of my published work on these issues, I have emphasized precisely this point, and I have discussed it at some length above. In my early work, I suggested that we shift our attention from worries about access to alternative scenarios to a more fine-grained analysis of the freedom-relevant properties of *the actual sequence*. Here I would simply suggest that, upon reflection, I believe that the case for the incompatibility of causal determinism and genuine access to alternative possibilities is considerably different from and *stronger than* the case for the incompatibility of causal determinism and sourcehood (in the sense linked to moral responsibility). A shift from considerations pertaining to alternative possibilities to sourcehood is thus dialectically felicitous for a compatibilist about causal determinism and moral responsibility.

# 7
# *Response to Kane, Fischer, and Vargas*

*Derk Pereboom*

## 1 Response to Kane

In his response, Robert Kane delivers a careful and sustained reply to three of the arguments against his view that I set out in chapter 3. First, Kane's reply to the Frankfurt-style case, Tax Evasion, is that the controller (or the device the controller has installed) "is not going to let Joe make the undetermined choice between A and B," where A is the choice to evade taxes, and B is doing otherwise, and thus Joe will not be (non-derivatively) morally responsible for the choice to evade taxes. If the cue for intervention, Joe's attaining the requisite level of attentiveness to moral reasons, does not occur, and he thus chooses A since the necessary condition for choosing B is not in place, then Joe's decision "will not be a 'will-setting' SFA (self-forming action) . . . because he will only have reasons to "set his will" on A and will not have attended to any good reasons to set his will on B." If he does attain the level of attentiveness, the controller will intervene and make him choose A. "So Joe will not get a chance to make a true SFA *either way* once the controller is in the picture."

Thus the reason Kane cites for Joe's not being non-derivatively morally responsible is that he will not have the undetermined choice between A and B. Notice that he is contending that Joe is not morally responsible because he cannot do otherwise. More precisely, Kane is claiming that Joe is not responsible because he lacks plural voluntary control, and in the sense specified by this notion, a robust alternative possibility. However, this is just the issue the leeway and the source theorist are arguing about, i.e., whether robust alternative possibilities are required for moral responsibility. In order to advance the debate, the source theorist devises a Frankfurt-style case in which

the agent lacks robust alternative possibilities, but which is intended to elicit the intuition that he is morally responsible. What are we then to say of the response that the agent is not responsible because he lacks robust alternative possibilities?

It would be mistaken to say that Kane's response begs the question against the Frankfurt-defender. For the success of a Frankfurt-style argument depends on whether the audience finds it intuitive that the agent is morally responsible. As it turns out, Kane does not find it intuitive that Joe is morally responsible. For him, the ultimate reason is that Joe lacks alternative possibilities, and this view may, in the last analysis, be correct. Notice, however, that there is a sense in which this is an extreme response to a Frankfurt-style case, since it cites the leeway position on what is at issue as the reason why Joe is not morally responsible. This response, in effect, precludes the possibility of discussing the issue at hand by way of Frankfurt-style cases. For we know in advance what, ultimately, the response to any such case will be: the agent is not responsible because he lacks robust alternative possibilities.

The Frankfurt-defender might hope that Kane's reaction to Tax Evasion will be unusual. It is not uncommon for people to have the intuition that in this example Joe is, or could be, morally responsible. For those who are undecided about whether he is morally responsible, let me again draw attention to the fact that the inevitability of Joe's decision is not grounded in its being produced by a causally deterministic process. For the absence of what would trigger the intervention at a particular time, that is, the absence of a certain level of attentiveness to moral reasons by a particular time, or a state indicated by this absence, does not, together with all the other actual facts about the situation, constitute a causally deterministic process that produces the decision. To see that this is so, take the intervener and his device from the scenario – this can be done legitimately, for by hypothesis, the device exerts no actual causal influence on Joe's deciding to evade taxes, so removing it will have no effect on whether he is causally determined to make this decision. There is no time at which refraining from deciding to evade taxes in the future is impossible for him, for he can always achieve the requisite level of attentiveness, and then he can freely refrain from deciding to evade taxes, or freely decide to evade taxes instead.

Kane's argument for requiring plural voluntary control is to be found in the "motives" part of his dual regress argument. There he contends that for an agent to set her will requires that she have access to what are in effect robust alternative possibilities. As Seth Shabo points out, Kane's concern here is whether the motivations present in a situation provide decisive reasons to choose as he does, and in a controversial case, one needs to ask whether the agent's will is set in this way (Shabo, unpublished manuscript). Kane's idea is that if the agent's will is set by the motivations present in the situation,

then non-derivative moral responsibility is precluded. Now one might argue that if there is no actual conflict of motivations for the agent, as is the case in Tax Evasion, then his will is set in the non-derivative responsibility-precluding way at issue. But this does not seem right. For even though there is no actual conflict of motivations for Joe, solely by way an exercise of his libertarian free will he could have been more attentive to the moral reasons, whereupon the conflict could have ensued. True, he would have had to be more attentive to these moral reasons than he actually was in order for them to motivate him, but nothing about the situation prevents him from achieving this level of attentiveness. So, intuitively, his will was not set in the non-derivative responsibility-precluding way by motivations present in the situation, or by anything else about the situation.

Let us now re-examine the luck objection to Kane's event-causal libertarianism. I contend that it allows us to see the problem for event-causal libertarianism most directly. In particular, this objection can be successful in showing why event-causal libertarianism cannot secure responsibility-conferring control. Consider Anne, the businesswoman, who can either decide to stop and help the assault victim, or can refrain from so deciding. The relevant causal conditions antecedent to this decision – agent-involving events, or, alternatively, states of the agent – would leave it open whether this decision will occur, and she has no further causal role in determining whether it does. I contend that with the causal role of the antecedent conditions already given, whether the decision occurs is not then settled by anything about the agent – whether it be states or events in which the agent is involved, or the agent herself. This fact provides a strong reason to conclude that the agent lacks the control required for being morally responsible for the decision.

Clarke points out that on the event-causal libertarian view, in addition to the agent's involvement in the antecedent events or states, there is a further respect in which the agent might be thought to contribute to a decision, and that is in *the causing of the decision* by the antecedent agent-involving states and events (2003: 74). The luck objection also challenges the supposition that this agent-involving causing of a decision provides for the agent's moral responsibility. Imagine that just prior to the occurrence of Anne's decision to stop, it was not determined whether refraining from deciding to stop be caused by one group of antecedent states R, or else deciding to stop be caused by another group S. On the event-causal libertarian picture, the causal conditions antecedent to the causing of the decision would leave it open whether the causing of Anne's refraining from deciding to stop by R, or else the causing of her decision to stop by S, will occur, and with the role of the antecedent conditions already given, the agent has no causal role in determining which of these options will occur. With the causal role of the antecedent events or states already in place, whether the causing of the decision by S

occurs is not settled by anything about the agent. This fact yields a significant reason to conclude that the agent lacks the control required for being morally responsible for the causing of the decision, and hence for the decision.

Kane argues that it is misleading to say that in this theory the agent has no further causal role to play in determining whether the decision occurs. For one thing, it suggests that "agents are not the causes at all of their undetermined self-forming actions or SFAs that all the agents do is somehow 'let them happen'." I do not object to the claim that under the envisioned event-causal libertarian circumstances, if the decision to stop is caused by S, it is also true that the agent, Anne, causes the decision to stop. My objection is not that on Kane's theory agents turn out not to be the causes of their SFAs, but rather that agents have no further role in the causing of their decisions than what the agent-involving antecedent conditions provide. On my own positive view, agents indeed have no further causal role than this in causing decisions, yet I do want to retain the claim that agents cause their actions, and that there is a sense in which they do not merely let them happen.

Kane argues that this objection leaves out the role efforts of will have in this theory. My concern about this key part of the view, as I argue in chapter 3, is that it is difficult to see how the agent can be morally responsible for a decision that results from an effort of will without being responsible for the effort of will itself. Given Kane's account of responsibility, it would seem that the agent cannot be responsible for an effort of will unless it is or results from a self-forming action. But if an agent can only be responsible for a self-forming action unless it results from an effort of will, it is difficult to see how responsibility can ever come about.

About the husband who swings his arm down on the table in an effort to break it, while it is not thereby causally determined whether the table will break, Kane says "if his effort succeeds, *he* (the agent) will have brought about the breaking of the table and he will be responsible for it." In his theory, efforts of will lie at the core of agency, in a way that, for example, desires do not. Suppose, by contrast, that an agent found herself with two desires, one to stop and help, the other to go on to the meeting, and that one of these desires would win out in the making of a decision, but it is not causally determined by antecedent states or events which wins out, and that these are all the key facts in the causal history of the decision. The agent would not seem to be morally responsible in this case. It is my sense that efforts of will do not differ from desires in this respect. Desires and efforts of will are simply states agents can be in or else agent-involving events, and whether it is desires or efforts of will that are in competition, only if the agent herself settles which of the competing states or events wins out to produce the decision can she be morally responsible for it. Perhaps only if the agent as substance has this settling power can this requirement be satisfied.

Finally, let us turn to the "no-further-power" objection. Clarke and I argue that moral responsibility demands control enhanced relative to what a deterministic account can provide, and that event-causal libertarianism cannot supply this enhanced control. Now Kane, Clarke, and I are incompatibilists, and thus we maintain that causal determinism precludes the sort of free will required for moral responsibility. I think that the best way to secure this claim is by the four-case manipulation argument, designed to show that there is no less reason to think that an agent's moral responsibility for a decision is ruled out by its being produced by a deterministic sequence of causal factors that traces back in time to factors beyond her control than there is to believe that it is ruled out by its resulting from a suitable kind of deterministic manipulation. As a result, responsibility requires control that is enhanced relative to what is possible in a deterministic context. But event-causal libertarianism, it seems, cannot provide this enhanced control. For if factors beyond the agent's control, rather than determining a single decision, instead simply leave it open which decision will occur, and the agent has no greater role in the production of this decision than she does in the deterministic context, then there is no more reason to think that she is morally responsible than there is in the deterministic context. So it appears that the event-causal libertarian can supplement the deterministic context only with relaxation of the causal net, and not with enhanced control.

But it remains open to the event-causal libertarian to suggest that causal determinism rules out control sufficient for moral responsibility only because it precludes alternative possibilities. Kane maintains, in effect, that the only reason determinism precludes this sort of control is that it rules out alternative possibilities of the robust sort, which are required for self-forming actions, without which moral responsibility would be impossible. But if a Frankfurt-style argument against the requirement of robust alternative possibilities for moral responsibility is successful, then it can't be that causal determination precludes responsibility simply because it rules out robust alternative possibilities. Then if alternative possibilities were required for moral responsibility, it would not be because an agent's having them *per se* would explain why she is morally responsible. Rather, alternative possibilities would be required indirectly, perhaps because the falsity of determinism was required for the agent to be the source of her action in a responsibility-conferring way, and the falsity of determinism demands alternative possibilities (Della Rocca 1998). So it may be that for agents to be morally responsible, they must have the power, as substances, to cause decisions without being causally determined to do so, and having this agent-causal power might require the availability of alternative possibilities. But supplementing the powers of deterministic agency with the availability of alternative possibilities will not all by itself make for the agent's being the source of her action in a responsibility-conferring way.

This is a good place to reply to a concern that Manuel Vargas raises for libertarianism. In chapter 4 he argues:

> . . . in that first instance of free will, and in every instance that follows, what control the agent has is a function of what options the world bestowed on that agent (through experience, heredity, socialization, circumstantial luck, and so on). Any control the agent has must be built up out of those constraints. Given that even the indeterministic options are thus constrained, and the elements that gave rise to those options . . . were not in control of the agent, what does the indeterminism give the agent in the way of control? Why doesn't the indeterminism simply open up multiple paths to an agent, where the constitution and sources of these paths were not something over which the agent had control?

The reply is this. Indeterminism is a necessary condition of responsibility-conferring control, but not simply because it supplies multiple paths. Only if indeterminism is true can agents, as substances, have the power to cause their actions without being causally determined to cause them, and this power might indeed be required for moral responsibility. If agents have this agent-causal power, multiple paths may well be open to them. But it is not simply by virtue of supplementing the powers of deterministic agency with multiple paths that control sufficient for moral responsibility is supplied – an agent might have these powers and multiple paths without having the agent-causal power.

## 2 Response to Fischer

In chapter 2, John Fischer presents several intuitive reasons to reject source incompatibilism, a position I defend. First, given that we are not troubled by the existence of causally necessary conditions for agency and for particular actions, such as the sun's shining on our planet, we should not be troubled by causally sufficient conditions for acting. Second, given that the source incompatibilist conditions on moral responsibility would be difficult or impossible to satisfy, it makes sense to articulate the notion of moral responsibility in such a way as to make our having it more likely.

In the last analysis, my response relies on the intuitive claim that if an action is produced by way of a deterministic process that traces back to causal factors beyond the agent's control, then he will not be morally responsible for it. If an agent decides to commit a crime, but this decision is produced in this deterministic way, it is strongly intuitive that he is not morally responsible for this decision. To be sure, one should regard this intuition just as a starting point, and as potentially defeasible. However, the four-case manipulation argument strengthens the force of this intuition, and my justification for

denying compatibilism is based on this argument. The source incompatibilist conditions on agency are indeed very difficult to meet. Perhaps we would need to be agent-causes – we would require the power, as substances, to cause decisions without being causally determined to cause them – and the likelihood that we have this power is not impressive. But it is the arguments that lead us to such a requirement, and I don't think that we should be strongly motivated to reject it simply by the unattractiveness of the resulting skepticism about moral responsibility. At the same time, I have tried to show that this position is not nearly as unattractive as it might initially appear to be.

Consequently, my incompatibilism is tied to the success of the four-case manipulation argument. As we saw in chapter 3, in Fischer's objection to this argument he contends that Professor Plum in Case 2 is morally responsible, but not blameworthy. In his response, he embellishes the claim that judgments of moral responsibility and judgments of blameworthiness and praiseworthiness are connected but distinct – "two separate but obviously related moments in our evaluation of behavior." Fischer's idea is that if an agent is legitimately judged morally responsible, he is then *eligible* to be judged blameworthy, but further considerations must be brought to bear before a judgment of blameworthiness is legitimate. I agree; once it settled that an agent is morally responsible, it then needs to be determined whether what he did was wrong, and whether he understood that it was wrong, and if he did not, whether he could have or should have understood that it was. But if it is settled that he is morally responsible for the action in the sense at issue in the debate, and that it was wrong and he understood that it was, then I claim that it is entailed that he is blameworthy for it.

I contend that, in general, an agent's being blameworthy for an action is entailed by his being morally responsible for it in the sense at issue in the debate, together with his understanding that the action was in fact morally wrong. This is because for an agent to be morally responsible for an action in the sense at issue is for it to belong to him in such a way that he would deserve blame if he understood that it was morally wrong, and he would deserve credit or perhaps praise if he understood that it was morally exemplary, supposing that this desert is basic in the sense that the agent would deserve the blame or credit just because he has performed the action, given understanding of its moral status, and not by virtue of consequentialist considerations. Assuming this characterization, and Plum's understanding that killing White is morally wrong, he could not be morally responsible for committing this murder without also being blameworthy for it.

Let me reiterate that there are other senses of "moral responsibility," and the thought that an agent be morally responsible for an action, understand that what he did was wrong, and yet not be blameworthy could be understood with reference to one of these senses. For example, an agent could be morally

responsible in the "legitimately called to moral improvement" sense; it would then be legitimate to expect him to respond to such questions as "Why did you decide to do that? Do you think it was the right thing to do?" and to evaluate critically what his decisions and actions indicate about his moral character. An intuition that Plum can be morally responsible without being blameworthy might be explained by his being responsible in this sense while not being blameworthy. But while this may be a *bona fide* notion of moral responsibility, it is not the one at issue in the free will debate. For incompatibilists would not find our being morally responsible in this sense to be even *prima facie* incompatible with determinism. The notion that incompatibilists do believe to be incompatible with determinism is rather the one defined in terms of basic desert, at least so I claim. If one wanted to pursue Fischer's strategy, one would need to specify a sense of moral responsibility that is plausibly the one at issue in the debate, and that allows for Plum not to be blameworthy while he is morally responsible and understands that his action is morally wrong.

The four-case argument serves to draw attention to the deterministic causes of action that would be present if determinism were true, but which would nonetheless typically be hidden from us. Spinoza observed, "men think themselves free, because they are conscious of their volitions and their appetite, and do not think, even in their dreams, of the causes by which they are disposed to wanting and willing, because they are ignorant of [those causes]" (Spinoza 1667/1985: 440). In his response, Fischer contends that we can make a distinction between two kinds of hidden causes, the first of which impairs responsibility, while the second does not. The first kind interferes with the normal functioning of mechanisms, while the second "is simply the set of constituents of the overt properties – these are the more specific or concrete ways in which the overt properties are instantiated." So, plausibly, our beliefs, desires, efforts of will, and decisions are instantiated in the neural structure of the brain. Fischer would contend that if the brain is functioning properly, the neural instantiation of properly reasons-responsive deliberation and action will not threaten our intuitive judgments of moral responsibility, even if the neural structure were governed by deterministic laws.

Fischer's key claim is that hidden causes of the second sort pose no threat to moral responsibility even if they are governed by deterministic laws. I disagree, and I base this judgment on the four-case argument, and, more generally, my case against compatibilism rests on the strength of the four-case argument. If an objection to this argument of the sort Fischer advances indicates that this argument is in fact unsound, then I would agree that hidden causes of the second sort fail to imperil moral responsibility.

All of that said, I think that the "legitimately called to moral improvement" sense of moral responsibility is perhaps the most significant one that

can be retained, and that it may well be that Fischer's notion of guidance control, developed in terms of reasons-responsiveness, yields the most promising account of this sort of responsibility. The ability to do otherwise would not explain how an agent might be morally responsible in this sense. What would explain it is the agent's capacity to appreciate the reasons for action that are present in the situation, and her ability to act in accord with these reasons. Causal determinism, and I suspect, any sort of indeterminism that is likely to be true, are compatible with our being morally responsible in this sense. So I suspect I can agree with Fischer on the following claim: his theory of moral responsibility – guidance control spelled out in terms of reasons-responsiveness – provides the most promising account of what might be the most significant sense of moral responsibility that can be retained given the best philosophical arguments and the best scientific theories we have about the physical world.

## 3 Response to Vargas

Vargas has us turn our attention to what in our ordinary conception of free will and moral responsibility should be revised, and what should be retained – a recommendation that yields a valuable perspective on the debate. His own diagnosis is that our ordinary conception of free will has a significant incompatibilist component, and his advice, in effect, is that this component should be revised so as to make it compatibilist. I would endorse this advice if at the same time our conception of moral responsibility, which has a major component the "basic desert" sense, were revised so as to eliminate it in favor of, say, the "legitimately called to moral improvement" notion. But perhaps Vargas's suggestion is to retain the "basic desert" sense of moral responsibility while revising our conception of the sort of free will required for it to a compatibilist one. Then his idea would be that while there are components of the ordinary view that are not friendly to compatibilism, we should eliminate them to become thoroughgoing and unconflicted compatibilists. I would counter with the "four-case" manipulation argument against compatibilism, and contend that if we revised our notion of free will to a compatibilist one, we would also need to revise our notion of moral responsibility so that the "basic desert" sense is eliminated.

In my view, the fundamental issue is whether our being morally responsible in the "basic desert" sense is compatible with determinism, or with the indeterministic ways the universe might well turn out to be. I think that it is compatible with neither, and that we are not morally responsible in the "basic desert" sense. This is the core of the hard incompatibilist view, as I conceive it. Nonetheless, we may still be morally responsible in, for example, the

"legitimately called to moral improvement" sense, and I think that in fact we are. I am very much open to the view that the question: "Are we sometimes morally responsible for our actions?" as posed in ordinary language, needs to be disambiguated. If it is specified that moral responsibility in the "legitimately called to moral improvement" sense is meant, then the answer is "yes." This answer would quite obviously not be inconsistent with hard incompatibilism, nor to incompatibilism more generally, as these notions functions in the debate. For what is at issue is whether moral responsibility in the "basic desert" sense is compatible with determinism (and with the relevant sorts of indeterminism).

"Free will," as I apply the concept in the context of the philosophical debate, designates the sort of free will required for moral responsibility in the "basic desert" sense, or, perhaps more precisely, whatever sort of control in deciding and acting is required for moral responsibility in this sense. This use of the concept "free will" is not intended to correspond with exactness to the use of the term "free will" in ordinary language. It is consistent with hard incompatibilism that there are correct applications of the ordinary language term "freely willed" to human actions. What this view claims is that human actions are never freely willed in the sense required for moral responsibility in the "basic desert" sense.

About the hard incompatibilist position, Vargas asks: "why think that we cannot change our thinking (or at least our theorizing) about free will in some way so as to render it less problematic?" Again, the arguments I advance for hard incompatibilism do not challenge the claim that some of our actions are freely willed in certain senses. Frankfurt, for example, specifies that a person *acts freely and of his own free will* just in case he wills X and wants to will X, and wills X *because* he wants to will X (1971). The arguments for hard incompatibilism pose no challenge to the claim that we sometimes act freely and of our own free will in Frankfurt's sense, and the hard incompatibilist position is consistent with this claim. It may well be that our ordinary concept "free will" and our ordinary thinking relevant to this concept allows for Frankfurt's use. So we might well not even need to change our thinking about free will to endorse, consistently, both hard incompatibilism and the claim that we in some sense act of our own free will. But if we would need to change our thinking about free will in order to do so, hard incompatibilism as I conceive it wouldn't necessarily oppose this move either.

Vargas contends: "what the arguments of hard incompatibilists never do, as far as I can tell, is show that there could not be a . . . disparity between our theoretical presuppositions about free will and the nature of free will itself. As long as there is this gap in the argument, we are not entitled to conclude that the implausibility of our self-conception is evidence that we are not free and responsible, for we might have free will but it might be different

than we tend to suppose." I think Vargas's concern turns out not to be a problem for hard incompatibilism, since the notion of free will that is at issue for this position builds the relevant theoretical presupposition into the notion itself: it is the sort of free will required for moral responsibility in the "basic desert" sense. Our lacking free will specified in this way is compatible with our agency having features to which the ordinary concept "free will" correctly applies.

Suppose we agreed that the concept "free will" now stands for just the sort of control required for moral responsibility in the "basic desert" sense. Imagine also that we endorsed Vargas's suggestion that we could revise the meaning of this concept, say by stipulation, and that we did revise it to stand for just the sort of control required for moral responsibility in the "legitimately called to moral improvement" sense. Vargas says: "the hard incompatibilist might claim that such freedom is not really free will." But given these suppositions, I would not contend that such freedom is not really free will. For the only concept in the area whose correct application to human decision and action I deny is "free will" in the sense required for moral responsibility as basic desert, and this allows for other concepts of free will to correctly apply.

Consider now Vargas's positive proposal for a moderate revisionism, which he characterizes as "incompatibilism about the diagnosis and compatibilism about the prescription . . . we might say that the account is incompatibilist about the folk concept of free will and compatibilist about what philosophical account we ought to have of free will." His account is revisionist in the sense that while there are incompatibilist elements in our ordinary beliefs and attitudes about free will, it recommends that we revise our concepts so as to eliminate these elements, and replace them with compatibilist alternatives. Given other claims Vargas makes, one might ask whether this is the most accurate characterization of his position. The key question is: does he endorse the view that some of our actions belong to us in such a way that we would, in the basic sense, deserve blame if we understood that they were morally wrong, and we would in this sense deserve credit or perhaps praise if we understood that they were morally exemplary? This is the notion of moral responsibility that incompatibilists think is incompatible with determinism, and compatibilists believe is compatible with determinism. When addressing the nature of moral responsibility, Vargas says, first, "the responsibility system aims to get creatures like us to better attend to what moral considerations there are and to appropriately govern our conduct in light of what moral reasons those considerations generate." So far, nothing has been said about responsibility that incompatibilists would believe to be incompatible with determinism. In fact what he goes on to say suggests the "legitimately called to moral improvement" notion: "over time, and given widespread participation in this system of judgements, practices, and attitudes, we come to help

both ourselves and other consideration-sensitive creatures to better track what moral considerations there are."

Fischer sets out the paradigmatic compatibilist position on the sort of responsibility we have, when in chapter 2 he remarks that someone who endorses his view "need not etiolate or reconfigure the widespread and natural idea that individuals *morally deserve* to be treated harshly in certain circumstances, and kindly in others. We need not in any way damp down our revulsion at heinous deeds, or our admiration for human goodness and even heroism." Where does Vargas stand on this crucial issue? He says: "instead of explicitly consequentialist responsibility norms, we should instead expect that the justified system of responsibility norms will look very familiar, accommodating both backwards-looking attitudes (such as gratitude) and forward-looking attitudes and practices." Here he appears to be endorsing the legitimacy of the backwards-looking reactive attitudes.

Now these backward-looking reactive attitudes presuppose beliefs to the effect that the agents to whom the attitudes are directed are deserving of those attitudes in the basic sense. Suppose, for example, that you read an account of a man committing a vicious murder (Watson 1987). As a result, you have a strong retributive attitude – you want the man to face the death penalty, and prefer that the death involve intense pain. However, you then learn that he was treated horribly in his upbringing, and you come to believe that his immoral character resulted from this treatment. At this point your retributive attitude diminishes, and perhaps disappears (although you might still believe that the killer is legitimately called to moral improvement). Arguably the best explanation for this change is that your retributive attitude presupposed the belief that the killer deserved, in the basic sense, to be the object of this attitude, and because you no longer have this belief, the attitude is deprived of the presupposition that sustained it. So then if Vargas aims to retain as justified backward-looking reactive attitudes, then it seems that he would have us believing that agents are morally responsible in the "basic desert" sense.

At this point two options are open to him. The first is to accept outright that agents are morally responsible in the "basic desert" sense, and to defend this claim against incompatibilist arguments, such as the four-case manipulation argument. Another is to deny that agents are morally responsible in the "basic desert" sense, but to argue that treating agents as if they are, and perhaps believing that agents are, achieves the best overall moral result (Pereboom 2001: 156). The first option is inconsistent with hard incompatibilism, but second is consistent with this position. I have contended that treating agents as if they are morally responsible in the "basic desert" sense might be justifiable on overall moral grounds, but that this option would have us believing and acting as if people deserve, in the basic sense, our

indignation, blame, and punishment when they really do not. But my sense is that supposing that hard incompatibilism is true, the better moral position would have us believe that agents are not morally responsible, and treat them as if they are not – insofar as we can. This is the view that I develop in chapter 3.

I thus advocate a position that in certain respects is more strongly revisionist about our attitudes and practices than the one Vargas appears to be defending. It would, for example, result in a radical change in how we assess moral resentment and indignation. We would attempt to avoid these attitudes where we can, perhaps substituting analogues, such as moral sadness. Our control over these reactive attitudes in the heat of an argument, or at the time one is personally wronged, is limited, and so in practice there might be little revision at this stage. Yet how we regard these attitudes afterwards would appropriately be subject to revision – we would think of them as lacking justification. Here we have significant control, and so revision is practically possible. We also have the relevant sort of control when we are deciding whether to adopt a system of criminal treatment that is justifiable only on retributivist grounds, or else one that can be justified in ways consistent with hard incompatibilism. At the same time, my position is not as decisively revisionist about the reactive attitudes required for good interpersonal relationships. Moral sadness, gratitude, forgiveness, and love are not undermined by hard incompatibilism, or else have analogues or aspects that would not have false presuppositions.

Is my view less revisionist about folk theory than Vargas's? If folk theory is incompatibilist in the sense that I specify, then in this respect my position is not revisionist at all, while Vargas's is. But what of the proposal that we come instead to think of free will as the control condition for moral responsibility in the "legitimately called to moral account" sense? That might be fine, but with a caveat. I suspect that the attitudes that presuppose basic desert – retributive sentiments, for example – will always be a feature of our psychology. I maintain that we need to retain concepts that allow us to think that these sentiments are irrational, and I would oppose any revisionism that proposes to dispense with these conceptual resources. More generally, when deciding how to revise, we need to retain concepts that facilitate our thinking that some of our attitudes and beliefs are mistaken. For this reason, we require concepts that apply to moral responsibility in the "basic desert" sense, and to the sort of control condition required for this sort of moral responsibility, even if these concepts do not correctly apply to anything real.

# 8

# *Response to Kane, Fischer, and Pereboom*

*Manuel Vargas*

Revisionism about free will is the view that there is a difference between what we think about free will and what we ought to think about free will. The revisionist account I favor is one that acknowledges that our commonsense thinking about agency has libertarian elements to it. In this respect, it agrees with libertarianism and disagrees with traditional compatibilism. Taken from the perspective of whether we should revise our current concepts, though, revisionism sides with hard incompatibilism in affirming that we should. Unlike hard incompatibilism, my account maintains that we need not abandon our deep notions of free will and moral responsibility. Instead, our concepts of free will and moral responsibility should be revised in a way that renders them compatible with the natural physical order, even a deterministic one. In this respect, revisionism agrees with compatibilism that the view we ought to have about free will is compatibilist. Thus, revisionism both overlaps with some aspects of each of the other alternatives and disagrees with others.

One artifact of this chapter's position in the book is that it benefits from being able to discuss the other response chapters. Partly, this is a matter of necessity; the only chapters to follow mine were response chapters, so there would be comparatively little for me to discuss if I did not address those chapters. Still, the other authors do not have a comparable opportunity to respond to my response, and undoubtedly they have more to say. So, it bears noting that this chapter is hardly the final word on these issues. I remain impressed by the other views, and, in particular, by their proponents in this book. Free will is an exceedingly difficult issue, and the other authors have offered compelling accounts that provide guidance on how to understand this thorny issue, even in the face of reasoned disagreement.

## 1 On Libertarianism, its Plausibility, and the State of Brain Science

Libertarianism's appeal is powerful, in part because it provides philosophical underpinnings for our pre-philosophical self-image. And, I am inclined to think that Robert Kane's account is one of the most plausible versions of libertarianism. Nevertheless, I am not a libertarian because of three families of considerations that weigh against it. One family of considerations is direct and two are indirect.

The first (and direct) objection is that libertarianism is implausible in the face of contemporary science. The second reason for skepticism about libertarianism concerns a puzzle about what the *normative* work is for the indeterminism requirement of libertarianism. If the indeterminism does no real normative work, then there is no special reason to hold a libertarian view, apart from preserving our self-image. Indeed, other things being equal, it seems better to have a theory less vulnerable to scientific rejection even if it isn't precisely what we had imagined from the start. If we were to have such a theory (and I think revisionist approaches have several good candidates), then libertarianism begins to look like an unnecessarily baroque philosophical theory, one that enshrines elements of our self-conception that were acquired in a pre-scientific era. A third and final element of my rejection of libertarianism is predicated on the idea that, from the standpoint of considerations about language and concepts, there is little reason to think that libertarianism is *uniquely* well suited to anchor our talk of free will. It may do the best job of capturing what we have in mind, but it is not strictly necessary for our sentences to turn out true. That is, there are plenty of good candidates for properties, or clusters, of properties, whose existence would make true most of our talk about responsibility, praise, and blame. If so, then it looks like we can do all the crucial work of a theory of free and responsible agency with a minimum of metaphysical calories.

In chapter 5, Kane says exactly the right things in response to my concerns about the scientific plausibility of libertarianism – indeed, I wish I had said those things. He is surely right that brain science theories are not explicitly deterministic, and that the scientific issues are not settled. In my experience, though, brain scientists tend to act as though their expectation is that our best accounts of the brain will be deterministic or close enough. That is, at least in conversation, libertarian accounts of agency tend to strike brain scientists as implausible. Kane is right to point out that these scientists could be in error. And, for what it is worth, I can imagine that I could be talked back into libertarianism if scientists found solid empirical data that showed that indeterminism is present in exactly all the right spots and few, if any, of the wrong ones. Since there is nothing that rules out the possibility that we will

make such a discovery, I agree with Kane that the scientific case against libertarianism is not decisive.

Even though Kane and I mostly agree about the state of the brain sciences we disagree about what conclusion to draw from it. In my view, it is striking that nothing in the brain sciences would lead an independent observer to conclude that we must be libertarian agents. If one is optimistic about the existence of libertarian agents, the optimism must come from somewhere else. If so, though, it would be good to know what evidence there is that supports libertarianism, given that the brain sciences are, at best, devoid of any real evidence for libertarianism.

Compare a belief in astral projection – perhaps it is *compatible* with the state of brain science, but compatibility does not mean plausibility, and I do not see why the fact that astral projection is not ruled out by brain science means that we should find astral projection plausible. If we are to find astral projection plausible, we would need some independent piece of evidence for its plausibility, and this is what we have yet to hear from proponents of astral projection. Similarly, what the libertarian needs is an account of *why* we should be optimistic that libertarianism will be vindicated by science. In contrast, I am inclined to think that the most reasonable thing for us to believe is that it is unlikely that our agency is of a sort described by Kane, but that we can get along just fine without it, anyway.

Kane does have a further reply available to him. It might be put something like this: since our self-image is libertarian, and since it is an open question whether science will bear out of self-image, we should proceed cautiously. The default assumption should be that our self-image is basically correct, and that assumption should only be abandoned it in the face of compelling evidence to the contrary. Since we lack such evidence, there is no compelling reason to yet abandon libertarianism. Although this reply might seem perfectly reasonable, I worry that such faith in our self-image comes at an unacceptable moral cost.

## 2 Welcome to the Jungle

Kane maintains that until we get compelling evidence, we should not forsake the Hawaii of libertarianism for the desert (of compatibilism?), the South Pole (of hard incompatibilism?) or the jungle (of revisionism?). While it is surely reasonable to want to spend some weeks in Hawaii while considering the alternatives, there are some good moral reasons for taking the first plane out of Hawaii.

Most libertarians will concede that there is no clear *proof* that we are libertarian agents. There might be some better and worse reasons to think we

are libertarian agents, but (supposing I am right) there is no clear evidence to suggest that we are indeed such agents. The best we are left with, then, is the hope that we are indeed libertarian agents. Consider, though, that some (though certainly not all) of our blame and punishment practices, in conception or in practice, presuppose libertarian free will. Most libertarians will also concede this point. Now, however, we face a problem. For any amount of blame and punishment based on the supposition that the agent has free will we are left in a sticky position with respect to that punishment. (Note, again, that we need not suppose that all blame or punishment presumes libertarianism. I am merely discussing the aspects of those things that do make such a supposition.) Consider what we can say to a skeptical subject of such blame and punishment, inquiring about why such punishment is justified. It seems that the libertarian can only say that such blame and punishment is justified if libertarianism is true, and that we hope – although we have no evidence to support such a hope – that the agent is, indeed, a libertarian agent. The subject of such blame and punishment need not dispute the *possibility* that he is, indeed a libertarian agent. Such a possibility is not, after all, incompatible with science. Nevertheless, such an agent is likely to protest that the mere possibility that he deserves such blame and punishment does not, by itself, make the punishment justified. After all, there is also a chance that he – and everyone else – might not be libertarian agents, at which point we would be faced with the unfortunate result that his blame and punishment are *un*justified. Since blame and punishment typically involve a kind of harm, there seems to be something morally objectionable about insisting on it based on the mere possibility that it could be justified. Or, to put the point a bit differently, the possibility of a Hawaiian vacation is cold comfort when sitting in a jail cell, or, worse, when facing execution. Thus, to the extent to which we sometimes blame and punish on putatively libertarian grounds (which, again, libertarians seem to think we do), we had better have a justification that runs deeper than the wish or hope that we are libertarian agents.

In contrast, my revisionism is in a better position than is libertarianism when it comes to the justification of any blame and punishment that hinges on free will. I have offered the outlines of an account of how we can justify our attributions of freedom and responsibility in ways that do not presume libertarianism, by appealing to the role that practices of praise and blame play in fostering a certain kind of intrinsically valuable agency, a kind of agency we already have good reason to believe exists. If the account works, then the revisionist is, at least with respect to the obligation to avoid unjustified harm, in a better position than the libertarian.

At this point the libertarian might simply take on board my kind of account of the justification of praise and blame. Or, the libertarian might treat

my account as a "Plan B" should libertarianism be proven mistaken. If so, then it looks like the libertarian is in no worse position than my kind of revisionist.

Here is the crucial point, though: once the libertarian helps him or herself to the sort of account I offer, the libertarian is in effect conceding that there is a libertarian-independent justification for our responsibility-characteristic practices. That is, the libertarian would be conceding that integrity of praise and blame does *not* depend on libertarian agency. If so, then what does libertarianism get you that you cannot get with a revisionist theory? If libertarians help themselves to revisionist resources, this threatens to undercut the impetus to be libertarian. To put the point a bit differently, once you visit the jungle, you might wonder if there was ever a reason to live anywhere else.

## 3 Desert and "Extra Factor" Accounts

At the end of chapter 1, Kane claims that agents are "to be conceived as *information-responsive complex dynamical systems*" (section 12). I think we can all agree with this characterization of our agency. What distinguishes our views is what else is required. For Kane, in addition to being an information-responsive complex dynamical system we also need indeterminism. I don't think we need indeterminism, but I do think we *want* it. I wonder, though, if some of us who are pre-philosophically (or folk) libertarians want something more than indeterminism. I sometimes wonder if part of what we want is a kind of independence from the causal order that no amount of well-located indeterminism can obtain. As Eddy Nahmias has emphasized in conversation, it is notable that the free will worry is sometimes expressed in terms of a threat to our agency in terms of our causal powers being built up out of lower-level natural phenomena (such as biological elements, brain chemistry, or even sub-atomic particles). This kind of threat to our free will might be called a *reductionist threat*. It is a threat that seems more rooted in our being built up out of smaller "material" things, than it is a fear about determinism, per se. For people who are motivated by the reductionist threat, it is hard to see how appeal to indeterminacies in the brain would salve the reductionist threat.

Recall that "extra factor" accounts are ones where agents are taken to have emergent and non-reducible causal powers, or where there is some other source of causal inputs, beyond the causal workings of low-level physical particles. There is a range of possible "extra factor" accounts, and what, precisely, these accounts come to is a very complicated thing and well beyond the scope of this chapter. Still, I sometimes wonder if the appeal of "extra factor" accounts (accounts which Kane rejects) is that they may suggest a way to answer both the deterministic threat and the reductionist threat. If one

can appeal to a non-reducible agential power, or a conception of irreducible agency, then the threat of reducing our free will to low-level phenomena diminishes.

It is unclear to me whether the reductionist threat really is a threat, and, if so, whether and how it can be answered. What is interesting in the present context is that anyone who thinks that an "extra factor" account of free will is the only way to adequately defend libertarianism will think that Kane's account is a kind of (perhaps inadvertent) revisionist libertarianism. But here Kane's incompatibilist critics must be careful. Because there is a range of "extra factor" views one might have, it is almost always possible that there will be some critic who objects that a given "extra factor" account fails to fully capture the power he or she believes is required for free will. Short of possessing what Roderick Chisholm once described as "godlike powers" of being akin to an unmoved mover, nearly any libertarian account may face a version of the "more libertarian than thou" objection.

Rather than attempting to capture every sincere conception of what might be required for free will, it seems to me that we are better off asking questions about what the presence or absence of various imagined forms of agency actually gets you or leaves you without. We can't always get what we want, but good enough should be just that – good enough. I suspect that Kane has no objection to these last two sentences. What he surely objects to is the idea that a non-libertarian account can secure what is crucial or essential to free will and moral responsibility. As he suggests in chapter 5, any non-libertarian account of free will strikes him as just another wretched subterfuge, substituting an ersatz notion of responsibility and freedom for the real things.

(However wretched my revisionism might be, I must emphasize that it cannot be a *subterfuge*. A subterfuge requires deception or a ruse in the implementation of a plan. On the contrary, the revisionist's plan is (I hope) transparently obvious. My approach does not pretend that the positive proposal is exactly what we had in mind all along, but I do contend that what it specifies is as good a candidate as we have for free will and moral responsibility.)

As I argued in chapter 4, the line of argument that insists on some incompatibilist element of free will being essential and unrevisable is not, thus far, compelling. We do not necessarily water down a concept just because we have revised it. Indeed, the objection suggests its own counterexample: Imagine a discussion with someone in the fourteenth century articulating a pre-chemical theory of water. It would strike us as unreasonable if such a person were to declare: "Either our pre-chemical theory of water will be vindicated by natural philosophy, or we will have watered down the meaning of water!" We should therefore accept that it is at least possible that one can change a concept without watering down the real meaning of the concept.

Still, one might insist that in this case no revision of our libertarian commonsense is possible without abandoning the concept. An argument would

be needed for this, and I have not thus far seen one. But even in the face of such an argument, I would be inclined to reply that revisionist free will is even better than the real thing, for on my view it has the comparative advantage of existing.

This brings me to some issues discussed in Pereboom's work. Pereboom emphasizes that he is interested in moral responsibility in the sense tied to deservingness (i.e., desert) of praise and blame. I agree that this is exactly the sense of moral responsibility with which we should be concerned. And, this is the sense of moral responsibility that, in my view, requires some degree of conceptual revision away from its libertarian commitments (and contrary to Pereboom's recommendation), *not* elimination. However, Pereboom goes on to claim that "The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just because she has performed the action (given an understanding of its moral status), and not by virtue of consequentialist considerations" (p. 86).

It is important to recognize that what Pereboom means by "basic desert" is a technical notion of desert, one that may or may not be the most fundamental moral notion of desert we use in connection with moral responsibility. Call this latter notion – the most fundamental moral notion of desert we use in connection with responsibility ascriptions, whatever it may be – *fundamental* desert. How should we understand fundamental desert? Pereboom's remarks may be taken to suggest that our notion of fundamental desert just is the notion of basic desert he describes. If so, this is a substantive view about fundamental desert, one that emphasizes the agent's performance of the action as the warrant for praise and blame. Still, one might maintain that basic desert is *not* fundamental desert, and that, as such, the existence or absence of basic desert is peripheral to debates about moral responsibility. To hold this view, however, one would need a substantive, non-basic conception of desert.

Basic and non-basic conceptions of fundamental desert could, at least in principle, agree on some things. For example, both basic and non-basic conceptions are compatible with the idea that moral properties or facts determine whether or not one deserves praise or blame. What distinguish non-basic views of fundamental desert are the particular moral properties or facts that underpin fundamental deservingness. A non-basic account might countenance, for example, facts about context, the social significance of the act, the role of the action in an economy of moral practices, the nature of our responses to said action, and anything that is not, roughly, a fact about the agent and his or her knowing performance of the act.

Outside of concerns for moral responsibility there are many non-basic senses of desert that are as fundamental as desert can be in those domains. For example, we might acknowledge that someone deserves fair opportunity

of employment, or the gold medal for running the fastest time at a meet, or a ticket for speeding. These notions of desert rely on a range of facts about human social practices, many of which are not basic in Pereboom's sense. Pereboom presumably agrees. What is at stake here is (1) whether the fundamental notion of deservingness implicated in responsibility assessments is basic in Pereboom's sense, and (2) if it is, whether it can be revised in a non-basic way. I do not yet see any reason why we should suppose that the fundamental sense of desert is basic, and I certainly do not see why we cannot revise it in some non-basic way if it is. For all that has been said in this volume, it does not seem that a real case has been made for either view. Until there is some argument on the table to favor one conception over the other, it seems to me that we should be open to the possibility that fundamental desert might have non-basic conditions.

On my account, there are good non-basic grounds for thinking that people can deserve praise and blame when they are responsible agents who have violated the norms of the responsibility system (see my discussion of the aims of the responsibility system, in chapter 4). We may *want* more than this, but it is not clear that we ever meant more. Nor is it clear what being deserving in the basic sense would allow us to do that we cannot do in a non-basic way, nor why it would be valuable to have that form of deservingness embedded in our practices. So, I would resist the idea that people cannot be genuinely deserving of moral praise and blame. Moreover, the grounds for our genuinely deserving moral praise and blame may not be basic, but it may well be as fundamental as our discourse of deservingness about moral responsibility can be.

## 4 The Four-case Argument

Pereboom's argument against compatibilist theories hinges on what he calls "the four-case argument." Since the prescriptive account I offer is intended to be compatible with determinism, it may be useful to remark on what a revisionist can say about the cases.

The aim of Pereboom's argument is to show that, for any agent who satisfies compatibilist conditions of free will and moral responsibility, there is no principled, responsibility-relevant distinction we can make between a case of direct neurological manipulation (which is presumably *not* a case of moral responsibility), and other cases of external causal influence, including determinism (pp. 93–8). If we think the first case – a case of direct neurological manipulation – is not a case of responsibility, then we should say the same thing about the other cases (including neurological and social "programming"), right up to the last case of ordinary determinism. Since we do think

that an agent is not morally responsible if he or she is directly neurologically manipulated, then we should say that if determinism is true, no one is responsible. And, if we say the latter, then we should say that compatibilists are mistaken: determinism and moral responsibility are incompatible.

I am inclined to think that Pereboom's "four-case argument" illustrates a prima facie problem for our commonsense conceptions of freedom and responsibility. However, I do not think it is especially problematic for the revisionist account I propose. Even if we concede that our pre-revisionist concept of free will provides us with few resources to make a principled distinction between the first and the fourth cases, there are principled ways a prescriptive revisionist can distinguish between the various cases. In this respect, I take my revisionist account to have an advantage, at least with respect to this argument, over ordinary libertarian and compatibilist accounts of free will.

It may be helpful to distinguish between two issues raised by the argument. One issue concerns whether or not the considered agent is a responsible *agent*, i.e., a candidate for evaluation in light of the norms of responsibility. The other issue concerns what the norms of responsibility (principally concerned with praise and blame) say about a particular agent in a particular circumstance, on the assumption that he or she is a responsible agent. This distinction is important to keep in mind because an agent may be appropriately subject to the norms of praise and blame (i.e., is a responsible agent), without those norms requiring that we praise or blame the agent. For example, it might well turn out that one is a responsible agent in some circumstance, but that the sort of action for which the agent is being evaluated is not one with moral significance one way or another. Alternately, an agent might be a responsible agent, and have done something morally significant, but the norms of praise and blame might recommend neither praising nor blaming the action. In each of these cases, we can say that the agent is a responsible agent without being either praiseworthy or blameworthy in that particular circumstance.

Although our terminology differs somewhat, John Fischer and I agree that the issue of agency and the issue of praise and blame can come apart (chapter 6). Pereboom does not make use of a similar distinction, so it is not completely clear to me whether his interest is in whether agents are responsible agents – i.e., candidates for praise and blame judgments, practices, and attitudes – or whether he is interested in what the norms of blame say about the considered agents in those circumstances, on the assumption that they are indeed responsible agents. I suspect that his concern is with the latter, but I will consider each possibility.

On the issue of responsible agency, at least cases 2–4 look like instances where the agent is (on my account) a responsible agent. That is, as long as

the agent really does have the basic structure of responsible agency (including the capacity to detect and appropriately respond to moral considerations), it strikes me as irrelevant, from the standpoint of the justified aims of a system of responsibility (and thus, my revisionism), whether and how the agent came to have those capacities. My hesitancy about the first case merely reflects my uncertainty about exactly what is involved in the neurological manipulation, and whether that manipulation really does leave the agent able to detect and appropriately respond to moral considerations. In principle, however, I am willing to concede that, depending on exactly how the neurological manipulation operates, it may be possible for an agent to retain his or her basic structure of responsible agency, and thus count as the kind of thing appropriately subject to the norms of praise and blame.

Concerning the application of the norms of praise and blame, things are more complicated. Recall that for blame to be appropriate, the agent must first be a responsible agent. Even if we grant that the agent in case 1 is a responsible agent, it does not follow that we should praise or blame the agent in that particular instance. To settle this issue, we must look to the norms of praise and blame. The blame norms get their content from several different sources, including the aims of a system of moral responsibility (on my account, fostering sensitivity to moral considerations), the limitations of our psychology and what we can reasonably demand of ourselves, and the permissions and requirements generated by the Right Theory of normative ethics – whatever that turns out to be. I have not attempted to defend a particular account of normative ethics, and so my theory of blame norms is necessarily indeterminate (which is not to say indeterministic!). Nevertheless, I do think there is reason to be doubtful that the norms of blame would permit blaming in the first case. How would blaming an agent subject to active neurological manipulation contribute to a stable and psychologically realistic system of judgments, practices, and attitudes directed at fostering moral consideration-responsive agency? We might be able to tell a story that makes sense of this, but it is hardly obvious how that story would go. Moreover, such a story would have to navigate around any further normative constraints imposed by the Right Theory of normative ethics. So, at least in the first case, blame looks misplaced on even a revisionist account.

In the second case, the "neurological programming" case, again, the details would matter. Pereboom is concerned with cases where the neurological programming is compatible with whatever the basic structure of responsible agency turns out to be, but we might imagine a case where the programming impaired or made more difficult the operation of those capacities without thereby eliminating them. In such a case, the agent may be blameworthy, but perhaps not fully so. If, through no endeavor or culpable failure of one's own, one's capacities are stretched beyond what we can reasonably expect of an

agent in that context, there is good reason to think that the force of blame should be diminished. This reflects the role of mitigation in our judgments of blameworthiness. We might think someone merits blame, or that we are permitted to blame, but facts about the agent in the context of action can also appropriately lead us to scale back the force or degree of blame. The reason for permitting mitigation in blame should be clear: it would be unreasonable for a system of responsibility to dismiss mitigation in cases where the reasonable demands on our capacity for morally sensitive, self-controlled behavior were exceeded.

Perhaps a natural question to ask is whether the best way to foster our moral considerations-sensitive agency is to have a system of strict and maximal blame on agents who have violated moral norms. Caution is in order, however. The emotional and social demands of always treating moral norm violations as fully blameworthy would be considerable, for everyone involved. Perhaps there are circumstances where mitigation practices have no benefits over less nuanced practices of blaming. Nevertheless, in many social arrangements, a mitigation-less pattern of human interaction would be incredibly taxing and disruptive. Moreover, the presence of mitigating practices might well come to be self-sustaining, to some degree. In contexts where the *de facto* norm is mitigated blame, unmitigated blame would constitute inequitable and unreasonable conduct, the sort of thing that itself would merit blame. Anyone who is a reader of this text is likely to live in a context with complex webs of moral and other interpersonal norms. In such contexts, insisting on systematic rejection of mitigation would likely be corrosive to our commitment to the responsibility system, and thus, the general development of moral consideration-sensitive agency. In our era, I doubt many would want to participate in a set of moral practices that were always and everywhere ruthless in their application of blame.

Mitigation, I maintain, is a plausibly justified aspect of our responsibility practices in the sociohistorical circumstances in which we find ourselves. Mitigation comes into play partly in response to our estimation of what it is reasonable to demand of one another. In light of the role that estimations of reasonableness play in mitigation, we might wonder how standards of reasonableness are to be settled. This is a thorny issue, and I am not committed to one or another particular view about how this issue might be settled. However, I do think it is plausible that our standards of reasonability will be partly indexed to what is familiar to us, the circumstances of our culture and moral upbringing, and so on.

For example, consider a belief in equality of opportunity. Believing that this form of equality ought to apply to people of non-European descent is comparatively easy for many people in our culture. It was much more difficult to believe this in the context of seventeenth-century Europe. Still, we should

think that most people in the seventeenth century were responsible agents. And, plausibly, inegalitarian attitudes are indeed blameworthy. Nevertheless, we can also admit that that the circumstances of their decision-making, the context in which they exercised their capacities for detection and response to moral considerations, were such that meeting the demands of morality on this particular issue for them was more difficult than it currently is for us.

What does any of this have to do with the four-case argument? Well, consider Pereboom's third case, where an agent has been determined by the training practices of his home and community. Although it is possible that on a revisionist account of responsibility we should recognize some degree of mitigation in judgments of blameworthiness, it is likely that a complete theory of moral responsibility, embedded in an account of normative ethics, will simply hold that the agent is fully blameworthy. Given the facts of the culture in which that agent is operating, and our estimation of what morality demands, the agent was not faced with unreasonable demands. Thus, the agent will, on this account, be fully blameworthy. Something similar is true of the last case as well (where the agent is determined by ordinary causal forces). On my view, such an agent is fully blameworthy.

So, I think that there are important differences between the various cases, and I think the revisionist can provide principled reasons to distinguish between these cases. My response *does not* presuppose that any of what I say is what we ordinarily think. Indeed, one benefit of reflecting on the four-case argument is that it provides a way of characterizing a difference between revisionism and hard incompatibilism; I think there is a principled way to distinguish between these cases and Pereboom does not.

## 5 Semicompatibilism and Revisionism: Will the Real Revisionist Please Stand Up?

On the variety of revisionism that I favor, we should revise our commonsense construal of the alternative possibilities requirement, any agent causation elements in our thinking, and, if we have them, any incompatibilist conception of a sourcehood requirement. In revising our concepts of free will and moral responsibility, we are *not* abandoning the idea that we are agents, that we have free will, or that we are morally responsible. What we are doing is reconceiving these things, but in ways that allow us to make principled and useful distinctions.

There is significant overlap between compatibilism and the kind of revisionism I recommend. On one way of looking at the issue, my revisionism can be considered a species of compatibilism. Philosophical labels tend towards plasticity, and the important thing is not the label but the

commitments of the theory. Thus, it is important to recognize substantial differences between revisionism and traditional conceptions of compatibilism. An obvious point of disagreement concerns the diagnosis of commonsense. Revisionism, at least my version of it, holds that commonsense has at least some incompatibilist elements. Compatibilists typically reject this claim. Perhaps more importantly, there are differences in the constraints faced by prescriptive revisionist theories on the one hand, and on the other, by traditional compatibilist theories. Revisionists are not bound by intuitions in the same way as compatibilists; revisionists are prepared to acknowledge a difference between what we believe and what we should believe and traditional compatibilists are not. For traditional compatibilists, if the theory gets the intuitions right, and if the theory provides some guidance on handling new or borderline cases, then it has done its work. Thus, traditional compatibilists face less immediate pressure to explain *why* the conditions of praise and blame should count as conditions for praise and blame. Revisionists, however, cannot always appeal to intuitions, for revisionists disavow those intuitions rooted in our putatively error-ridden folk concepts. Consequently, the revisionist faces a stronger demand to explain why the conditions for praise or blame are as he or she proposes.

In light of these differences, it is striking that in chapter 6, John Fischer writes, "semicompatibilism is a significant revision of ordinary, commonsense thinking – as well as standard philosophical reflection on freedom and moral responsibility" (p. 188). He goes on to say that he does not "deny that alternative possibilities are a presupposition of commonsense as well as philosophical analysis" and that his account of the conditions of moral responsibility "are significantly revisionary (perhaps even revolutionary)" (p. 189).

I would love to have such distinguished support for the view I favor, but I doubt that Fischer is a revisionist in my sense. To see why, it helps to distinguish between weak, moderate, and strong revisionisms. To foreshadow: my discussion of "revisionism" has presumed moderate revisionism, and I doubt Fischer is a moderate revisionist.

Strong revisionism is essentially a (metaphysically) skeptical view. It holds that the correct prescriptive account is one that jettisons talk of responsibility and free will, at least in the senses that are central to free will debates. For example, Pereboom's account is revisionist in this sense. Since "hard incompatibilism" is a perfectly good label for a substantive position, though, and since there is good reason to treat that kind of revisionism as a *sui generis* category, it makes sense to count hard incompatibilism as distinct from the sort of revisionism here at stake. It is also clear that Fischer does not mean his account is revisionist in the strong sense that applies to Pereboom's view.

This leaves us with two other forms of revisionism: weak and moderate. To see the difference between weak and moderate revisionism, reflect on the

difference between, on the one hand, what the folk think, and on the other hand, what the folk *think* they think. *Weak* revisionism is revisionism about what the folk think they think; it is the idea that the folk have in some way failed to appreciate the nature of their own conceptual or metaphysical commitments. While the folk really believe X, the folk mistakenly understand themselves to believe Y. In contrast, *moderate* revisionism is revisionism about what the folk think.

For my purposes, the paradigmatic variety of revisionism – the revisionism I have been concerned to defend – is moderate revisionism. It is what I mean when I talk about revisionist approaches to responsibility. If Fischer is a kind of revisionist, he is a weak revisionist. Or he ought to be, on pain of inconsistency.

Suppose that we wanted to construe Fischer as a moderate revisionist. To see why this would be problematic, consider his remarks on alternative possibilities. Fischer claims that commonsense presupposes alternative possibilities (p. 188). However, as he makes clear in this volume and in numerous other places, he also thinks Frankfurt-cases are illuminating. How could that be? The whole point of Frankfurt-cases is to illustrate that despite what we may have thought, responsibility ascriptions do *not* presuppose the existence of alternative possibilities. In other words, what we learn from Frankfurt-cases is that alternative possibilities are not required by our present concept of responsibility. So, if you accept that Frankfurt-cases are genuinely informative (which Fischer clearly does), then you have to think that our commonsense concept of responsibility does *not* include an alternative possibilities requirement.

(Another way to put the point that Fischer takes from Frankfurt-cases is to say that alternative possibilities are not part of the conditions of application for the concept of moral responsibility. Fischer draws a sharp distinction between concepts and their conditions of application. I am skeptical about this distinction, as one might plausibly think that the conditions of application are simply part of a concept, and that attempts to draw finer-grained distinctions are bound to fail, or, at any rate, are not representative of our best accounts of the nature of concepts. Still, I think that the essential difference between weak and moderate revisionism can be generated at both the level of the concept and the level of conditions of application for the concept.)

It thus looks like Fischer can't believe – as he claims to – that commonsense *does* presuppose alternative possibilities. It looks like Fischer either has to give up his (moderate) revisionism, or his allegiance to the apparent lesson of Frankfurt-cases.

However, if Fischer is a moderate revisionist, then this is a substantial and new concession. It would have consequences for many of the positions he has long held in print. As I argue above, it would minimally mean that many of his former defenses of the lesson of Frankfurt-cases would have to be

abandoned. Moreover, his concern for a historical condition on moral responsibility would be puzzling. Why care about the history of how a mechanism is acquired, if one is prepared to be a revisionist? Indeed, if Fischer is prepared to embrace moderate revisionism, then there is no reason why he cannot dismiss Pereboom's four-case argument by acknowledging that his positive account of free will is not intuitive. Indeed, if Fischer were a moderate revisionist, his approach to a wide range of philosophical intuition pumps becomes puzzling; answering a range of traditional philosophical arguments becomes much less pressing if one is prepared to walk away from folk intuitions.

Since nothing Fischer writes suggests he is prepared to abandon his views on these issues, it seems we are better off interpreting Fischer as a weak revisionist, and not as a moderate revisionist. On this interpretation of his view, his declarations of revisionism amount to the claim that what the folk think they think is that alternative possibilities are required for moral responsibility. What Frankfurt cases show is that the folk are mistaken about what they think they think. So, in light of Frankfurt cases, we should revise what we think we think.

Interestingly, many traditional compatibilist accounts have been weakly revisionist. Conditional analysis-style compatibilists oftentimes maintained that if we just properly understood the meanings of our words we would realize that we were never committed to a conception of moral responsibility that was incompatible with determinism. In this respect, semicompatibilism is in the same boat with traditional compatibilism. Both traditional compatibilists and semicompatibilists maintain that the folk are mistaken in what they think they think, and both maintain that what we *really* think is compatible with determinism. Moreover, both can and oftentimes do concede that there is some sense of ability that would be ruled out by determinism. Crucially, though, both hold that the sense ruled out be determinism is not the one required for moral responsibility. When you combine this point with the traditional idea that free will just is the freedom or control condition on moral responsibility, the gap between semicompatibilism and classical compatibilism becomes very compressed.

Of course, I hope I am wrong about Fischer. By my lights, it would be wonderful if he really converted to (moderate) revisionism, and, thus, felt less concern for a range of intuitions that he has been largely concerned to preserve. But I suspect he hasn't – yet.

## 6 The Debate That Does Not End

For better or for worse, all books must end, even when disagreements contained in it do not.

Sometimes this latter fact – the seemingly indefatigable persistence of philosophical debates – is itself taken to be problematic. Indeed, sometimes recognition that philosophical debate seems interminable can give rise to the view that in places where disagreement persists, no resolution is possible, even in principle (e.g., van Inwagen 1996; McGinn 1993).

Perhaps there are contexts in which we should draw this conclusion, but I am skeptical this is one of them. Few, if any, fields of human inquiry have anything approaching full convergence. Still, one might think things are particularly dire in the context of reflection on free will. One might be tempted to conclude that even if in principle progress could be made, our theorizing is not yet up to the demands of the topic, for we have been grappling with free will for a very long time and yet we still disagree about fundamental issues.

We should resist this conclusion. I think it reflects an underappreciation of what progress has been made in, say, the last 40 years of work on the free will problem. We are, for example, much clearer about what the options are for all the major positions. We have a deeper appreciation of the complexity of things like the principle of alternative possibilities. More generally there is, I think, a growing appreciation of the sheer number of philosophical puzzles entangled in the free will debate. Recognizing the complexity problem is a kind of progress and, plausibly, a condition for the possibility of robust progress (or, at least, more substantial convergence).

More controversially, I am inclined to think that there *is* growing consensus about those approaches to free will issues that are promising and those that are not. For example, few compatibilists today think the free will problem is resolved by appealing to a simple conditional analysis of "can," and considerably fewer libertarians seem comfortable appealing to non-natural properties in their account of free agency. Forty years ago, there was nothing like the degree of convergence we now have about these things.

To be sure, debate – even among philosophers – seldom convinces the interlocutors to change their minds about anything substantial, at least not right away. Even if full resolution of these issues among philosophers remains elusive, the prospect for a different kind of success seems good: a debate may also succeed is if it encourages others to contribute to a fruitful discussion.

# *Bibliography*


Anglin, W.S. (1991) *Free Will and the Christian Faith* (Oxford: Oxford University Press).

Anscombe, G.E.M. (1971) *Causality and Determinism* (Cambridge: Cambridge University Press).

Austin, J.L. (1966) "Ifs and Cans," in Bernard Berofsky (ed.), *Free Will and Determinism* (New York: Harper and Row), pp. 295–321.

Balaguer, Mark. (n.d.) *Is Free Will an Open Scientific Question?* Unpublished manuscript, chapter 4.

Bok, Hilary (1998) *Freedom and Responsibility* (Princeton: Princeton University Press).

Chisholm, Roderick (1976) *Person and Object* (La Salle, IL: Open Court, 1976).

Clarke, Randolph (1994) "Toward a Credible Agent-Causal Account of Free Will," *Noûs* 27, 191–203.

Clarke, Randolph (1995) "Indeterminism and Control," *American Philosophical Quarterly* 125–38.

Clarke, Randolph (1997) "On the Possibility of Rational Free Action," *Philosophical Studies* 88, pp. 37–57.

Clarke, Randolph (2000) "Modest Libertarianism," *Philosophical Perspectives* 14, 21–45.

Clarke, Randolph (2003) *Libertarian Theories of Free Will* (Oxford: Oxford University Press).

Dayan, P. and L.F. Abbott (2001) *Theoretical Neuroscience* (Cambridge, MA: MIT Press).

Della Rocca, Michael (1998) "Frankfurt, Fischer, and Flickers," *Noûs* 32, 99–105.

Ekstrom, Laura W. (2000) *Free Will: A Philosophical Study* (Boulder, CO: Westview Press, 2000).

Farrell, Daniel M. (1985) "The Justification of General Deterrence," *The Philosophical Review* 104 (1985), 367–94.

Fischer, John Martin (1982) "Responsibility and Control," *Journal of Philosophy* 79, 24–40.

Fischer, John Martin (1994) *The Metaphysics of Free Will* (Oxford: Blackwell).

Fischer, John Martin (1999) "Recent Work on Moral Responsibility," *Ethics* 110, 93–139.

Fischer, John Martin (2004) "Responsibility and Manipulation," *The Journal of Ethics* 8, 145–77.

Fischer, John Martin, and Mark Ravizza (1998) *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press).

Foot, Philippa (1957) "Free Will as Involving Determinism," *The Philosophical Review* 66, 439–50.

Frankfurt, Harry G. (1969) "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66, 829–39.

Frankfurt, Harry G. (1971) "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, 5–20.

Ginet, Carl (1996) "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Arguments Convincing," *Philosophical Perspectives* 10, 403–17.

Ginet, Carl (1997) "Freedom, Responsibility, and Agency," *Journal of Ethics* 1, 85–98.

Goetz, Stewart (2002) "Alternative Frankfurt-Style Counterexamples to the Principle of Alternative Possibilities," *Pacific Philosophical Quarterly* 83, 131–47.

Haji, Ishtiyaque (1998) *Moral Appraisability* (New York: Oxford University Press).

Haji, Ishtiyaque (2004) "Active Control, Agent Causation, and Free Action," *Philosophical Explorations* 7, 131–48.

Honderich, Ted (1988) *A Theory of Determinism* (Oxford: Oxford University Press).

Hume, David (1739/1978) *A Treatise of Human Nature*, ed. L.A. Selby-Bigge (Oxford: Oxford University Press).

Hunt, David (2000) "Moral Responsibility and Unavoidable Action," *Philosophical Studies* 97, 195–227.

Hunt, David (2005) "Moral Responsibility and Buffered Alternatives," *Midwest Studies in Philosophy* 29, 126–45.

Kane, Robert (1985) *Free Will and Values* (Albany: SUNY Press).

Kane, Robert (1996) *The Significance of Free Will* (New York: Oxford University Press).

Kane, Robert (1999) "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism," *Journal of Philosophy* 96, 217–40.

Kane, Robert (ed.) (2002) *The Oxford Handbook of Free Will* (New York: Oxford University Press).

Lycan, William G. (1997) *Consciousness* (Cambridge, MA: MIT Press).

McGinn, Colin (1993) *Problems in Philosophy: The Limits of Inquiry* (Cambridge, MA: Blackwell).

McKenna, Michael (1997) "Alternative Possibilities and the Failure of the Counterexample Strategy," *Journal of Social Philosophy* 28, 71–85.

McKenna, Michael (2003) "Robustness, Control, and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives," in D. Widerker and M. McKenna (eds.), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Aldershot: Ashgate Press).

McKenna, Michael (2005a) "Where Frankfurt and Strawson Meet," *Midwest Studies in Philosophy* 29, 163–80.

McKenna, Michael (2005b) "The Relationship Between Autonomous and Morally Responsible Agency," in J.S. Taylor (ed.), *Personal Autonomy* (Cambridge: Cambridge University Press), pp. 205–34.

Mele, Alfred (1999) "Ultimate Responsibility and Dumb Luck," *Social Philosophy and Policy* 16, 274–93.

Mele, Alfred (2005) "Libertarianism, Luck, and Control," *Pacific Philosophical Quarterly* 86, 395–421.

Mele, Alfred (2006) *Free Will and Luck* (Oxford: Oxford University Press).

Nagel, Thomas (1979) "Moral Luck," in *Mortal Questions* (Cambridge: Cambridge University Press), pp. 24–38.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner (forthcoming) "Is Incompatibilism Intuitive?," *Philosophy and Phenomenological Research*.

Nichols, Shaun (2006) "Folk Intuitions on Free Will," *Journal of Cognition and Culture*, 6(1&2), 57–86.

Nichols, Shaun and Joshua Knobe (forthcoming) "Moral Responsibility and Determinism: The Cognitive Science of Folk Institutions," *Noûs*.

Nietzsche, Friedrich Wilhelm (1971) *On the Genealogy of Morality* (Indianapolis, IN: Hackett).

Nietzsche, Friedrich Wilhelm (1982) *Daybreak: Thought on the Prejudices of Morality* (Cambridge: Cambridge Univesity Press).

Nietzsche, Friedrich Wilhelm (1996) *On the Genealogy of Morality* (Indianapolis, IN: Hackett).

O'Connor, Timothy (2000) *Persons and Causes* (Oxford: Oxford University Press).

O'Connor, Timothy (2003) "Review of *Living Without Free Will*," *Philosophical Quarterly* 53, 308–10.

Otsuka, Michael (1998) "Incompatibilism and the Avoidability of Blame," *Ethics* 108, 685–701.

Pereboom, Derk (1995) "Determinism *Al Dente*," *Noûs* 29, 21–45.

Pereboom, Derk (2000) "Alternative Possibilities and Causal Histories," *Philosophical Perspectives* 14, 119–37.

Pereboom, Derk (2001) *Living Without Free Will* (Cambridge: Cambridge University Press).

Pereboom, Derk (2003) "Source Incompatibilism and Alternative Possibilities," in Michael McKenna and David Widerker (eds.), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Aldershot: Ashgate), pp. 185–99.

Pereboom, Derk (2004) "Is Our Conception of Agent-Causation Coherent?," *Philosophical Topics* 32, 275–86.

Pereboom, Derk (2005) "Defending Hard Incompatibilism," *Midwest Studies in Philosophy* 29, 228–47.

Rawls, John (1971) *A Theory of Justice* (Cambridge, MA: Harvard University Press).

Schoeman, Ferdinand D. (1979) "On Incapacitating the Dangerous," *American Philosophical Quarterly* 16, 27–35.

Shabo, Seth (manuscript) "Uncompromising Source Incompabilism."

Smilansky, Saul (1997) "Can a Determinist Help Herself?," in C.H. Manekin and M. Kellner (eds.), *Freedom and Moral Responsibility: General and Jewish Perspectives* (College Park: University of Maryland Press), pp. 85–98.

Smilansky, Saul (2000) *Free Will and Illusion* (Oxford: Oxford University Press).

Spinoza, Baruch (1677/1985) *Ethics*, in *The Collected Works of Spinoza*, ed. and tr. Edwin Curley, volume 1 (Princeton: Princeton University Press).

Strawson, Galen (1986) *Freedom and Belief* (Oxford: Oxford University Press).

Strawson, P.F. (1962) "Freedom and Resentment," *Proceedings of the British Academy* 48, 1–25.

Strawson, P.F. (1992) *Analysis and Metaphysics* (New York: Oxford University Press).

van Inwagen, Peter (1983) *An Essay On Free Will* (Oxford: Oxford University Press).

van Inwagen, Peter (1996) "Review of *Problems in* Philosophy," 105(2), 253–6.

Vargas, Manuel (2005) "The Revisionist's Guide to Moral Responsibility," *Philosophical Studies* 125, 399–429.

Wallace, R. Jay (1994) *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press).

Waller, Bruce (1990) *Freedom Without Responsibility* (Philadelphia: Temple University Press).

Walter, Henrik (2001) *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy* (Cambridge, MA: MIT Press).

Watson, Gary (1987) "Responsibility and the Limits of Evil," in Ferdinand Schoeman (ed.), *Responsibility, Character, and the Emotions* (Cambridge: Cambridge University Press), pp. 256–86.

Widerker, David (1995) "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *The Philosophical Review* 104, 247–61.

Widerker, David (2000) "Frankfurt's Attack on Alternative Possibilities," *Philosophical Perspectives* 14, 181–201.

Woolfolk, Robert L., John Doris, and John Darley (2006) "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility," *Cognition* 100(2), 283–301.

Wyma, Keith D. (1997) "Moral Responsibility and Leeway for Action," *American Philosophical Quarterly* 34, 57–70.

# *Index*