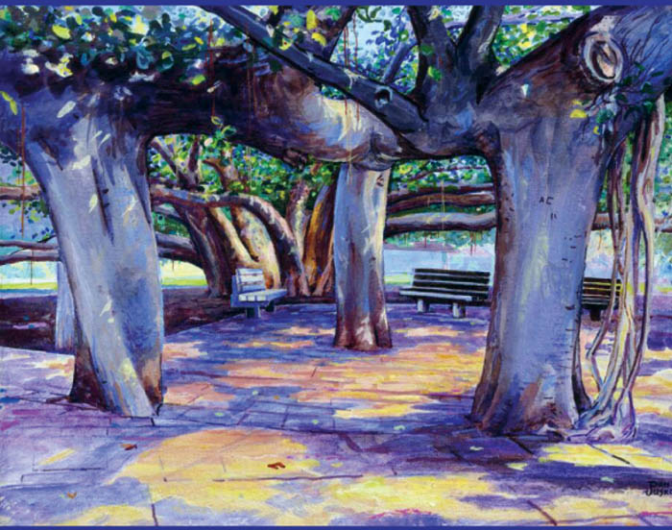
Beyond Reduction

Philosophy of Mind and Post-Reductionist Philosophy of Science



Steven Horst

Beyond Reduction

PHILOSOPHY OF MIND

SERIES EDITOR
David J. Chalmers, Australian National University

Self Expressions
Minds, Morals, and the Meaning of Life
Owen Flanagan

The Conscious Mind
In Search of a Fundamental Theory
David J. Chalmers

Deconstructing the Mind
Stephen P. Stich

The Human Animal
Personal Identity without Psychology
Eric Olson

Minds and Bodies
Philosophers and Their Ideas
Colin McGinn

What's Within?
Nativism Reconsidered
Fiona Cowie

Dreaming Souls
Sleep, Dreams, and the Evolution of the Conscious Mind
Owen Flanagan

Purple Haze
The Puzzle of Consciousness
Joseph Levine

Consciousness and Cognition
A Unified Account
Michael Thau

Thinking without Words
José Luis Bermúdez

Identifying the Mind
Selected Papers of U. T. Place
Edited by George Graham and Elizabeth R. Valentine

A Place for Consciousness
Probing the Deep Structure of the Natural World
Gregg Rosenberg

Three Faces of Desire
Timothy Schroder

Gut Reactions
A Perceptual Theory of Emotion
Jesse J. Prinz

Ignorance and Imagination
On the Epistemic Origin of the Problem of Consciousness
Daniel Stoljar

Simulating Minds
The Philosophy, Psychology, and Neuroscience of Mindreading
Alvin I. Goldman

Phenomenal Concepts and Phenomenal Knowledge
New Essays on Consciousness and Physicalism
Edited by Torin Alter and Sven Walter

Beyond Reduction

Philosophy of Mind and Post-Reductionist Philosophy of Science

STEVEN HORST

OXFORD
UNIVERSITY PRESS
2007

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Horst, Steven W., 1960–
Beyond reduction: philosophy of mind and post-reductionist
philosophy of science/Steven Horst.
p. cm.—(Philosophy of mind)
Includes bibliographical references and index.
ISBN 978-0-19-531711-4
1. Philosophy of mind. 2. Reductionism. I. Title.
BD418.3.H669 2007
128′.3—dc22 2006052764

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

For my mother

This page intentionally left blank

Preface

It is difficult to fix a date for the beginnings of this book with any real precision. It is one of several fruits of a project that began in the summer of 1993. At that time, having finished all but editorial work on *Symbols, Computation and Intentionality* (Horst 1996), I began to explore a topic on which I had left a substantial promissory note in that book: namely, the question of whether mental phenomena like intentionality and phenomenology could be naturalized. It was with this in mind that I attended NEH Summer Institutes on Naturalism (at the University of Nebraska, hosted by Robert Audi) and Meaning (at Rutgers University, hosted by Jerry Fodor and Ernie LePore). One thing that became clear to everyone at Audi's institute was that, while a great number of philosophers wish to lay claim to the *word* 'naturalism', they in fact use that word in a surprising number of ways. Chapter 1 of this book, which attempts to bring some order to this motley assortment of usages, grew out of extended research into the contemporary and historical usages of the term and the research projects associated with it.

When I started out on the project, I still assumed, as I had in *Symbols, Computation and Intentionality*, that intertheoretic reductions were the rule in the natural sciences and that the explanatory gaps encountered with respect to consciousness, intentionality, and normativity present unique problems. During a 1997–98 sabbatical at Princeton and Stanford's Center for the Study of Language and Information, made possible by an NEH Fellowship and by sabbatical support from my home institution, Wesleyan University, I had several conversations with philosophers of science (Paul Humphreys, Bas van Fraassen, Patrick Suppes, and my Wesleyan colleague Joseph Rouse) who regarded my

reductionist assumptions in the philosophy of the natural sciences with some incredulity and pointed out to me that the kind of reductionism I was assuming had been largely rejected (and for good reasons) within philosophy of science itself. This led to a gradual transformation of how I viewed my project, and to an overarching question that I would now put like this in its most general form: *What ought philosophy of mind to learn from contemporary philosophy of science?* (And, in its more pointed form, *Why is contemporary philosophy of mind one of the last bastions of the philosophy of science of the 1950s?)* Both this book, and several other descendents of the larger project (Horst 2004, 2005) attempt to address various aspects of the general question.

Initially, I conceived of this project as one book, to be published under the title *Mind and the World of Nature*, that would explore not only the variety and prospects of contemporary naturalistic approaches to the mind, but also their historical roots in particular views of scientific explanation dating back to around 1600 and case studies in explanation in the sciences of the mind that would explore whether there do indeed seem to be real and abiding explanatory gaps even after decades of research in psychology and neuroscience. As a reasonable person might have expected, the resulting manuscript grew to unworkable proportions. I am indebted to the various people who read all or parts of the initial nine-hundred-page version of that manuscript, whose efforts must have been truly heroic. These include Carol Slater of Alma College in Michigan (who supplied copious comments in a lovely purple ink from what I suspect to be one of those by-golly fountain pens that I have never been able to master), Eric Schwitzgebel of the University of California at Riverside and the intrepid band of graduate students in his seminar (who not only read the manuscript but grilled me on it during a visit to Riverside where they kindly put me up in the Mission Inn, perhaps the only hotel I have ever visited that I would be tempted to go back to just to experience the hotel itself again), and a very fine group of Wesleyan undergraduates in my Topics in Philosophy of Mind seminar that began in the tempestuous month of September 2001. A slightly trimmer version was read by two anonymous referees, who confirmed my suspicions that this was really not a single book but several books, with different topics, for different audiences, while also providing helpful (and sympathetic) feedback on many of the main points.

The book you are now reading, while descended from those drafts of *Mind and the World of Nature*, involved a complete rewriting of everything, with a narrower orientation and greater focus upon reductive forms of naturalism. In its final form, it is particularly indebted to suggestions from Thomas Polger of the University of Cincinnati, who read the penultimate draft. Tom's own book (Polger 2003) defends a form of type-identity theory, on which account he might seem to represent a point of view almost completely antithetical to my own view, which is not only antireductionist but antinaturalist. However, he and I actually agree on a number of points, ranging from the failure of the classic reductionist project of Carnap and Nagel in philosophy of science to the need to adopt some form of pluralism in philosophy of science and philosophy of mind. However, the forms our pluralisms take are quite different. Thanks

also go to a number of people who kindly agreed to read chapters at various stages along the way, particularly Michael Silberstein and Joe Rouse. The final product is better for their input.

Neither memory nor space allows me to give due credit to all of the people with whom I have had profitable conversations over the years that have helped to shape the final form of this book. I shall single out a few, hopefully without giving offense to those who have been omitted. Wesleyan University has supported this work through two semester-long sabbaticals. David Chalmers, now at the Australian National University, has been a continuing source of lively engagement as our views have drifted apart over the past decade, and was so kind as to include me in the NEH Summer Institute on Consciousness and Intentionality he and David Hoy hosted at the University of California at Santa Cruz in 1992. My Wesleyan colleague Joseph Rouse has been helpful on a great number of occasions in pointing me in useful directions in philosophy of science. Much of the book's crucial turn toward post-reductionist philosophy of science might never have come about without his ever-generous colleagueship. My Wesleyan colleague Sanford Shieh has on several occasions provided insights (which I have probably not adequately appropriated) on modal logic and its applications to metaphysics. Michael Silberstein (Elizabethtown College and University of Maryland), William Bechtel (University of California, San Diego), John Bickle (University of Cincinnati), Paul Churchland (University of California, San Diego), Peter Godfrey-Smith (Harvard and the Australian National University), Jaegwon Kim (Brown University), and Abner Shimony (Boston University) have been fine interlocutors on the question of what philosophy of mind might learn from contemporary philosophy of science, at forums sponsored by the Society for Philosophy and Psychology and the Boston Center for the Philosophy and History of Science, hosted by Alfred Tauber and Robert Cohen, whom I regard as the Father Mersenne of late twentieth-century philosophy of science. I hope that such conversations will help usher in a new era in philosophy of mind, one better engaged with both philosophy of science and the details of the various sciences of cognition. Thanks also to Stanford's Center for the Study of Language and Information and to the philosophy departments at the University of California at San Diego, University of Connecticut, Calvin College, and Elizabethtown College, which hosted talks at which I presented versions of material contained herein. Special thanks go to the editors at Oxford University Press, first Robert Miller and then Peter Ohlin, and to the series editors who solicited the manuscript, first Owen Flanagan and then David Chalmers. And I am greatly indebted to Doretta Wildes, who proofread the book. The final version of the book is substantially better for the suggestions of three referees contracted by Oxford University Press. The full list is, of course, much longer. To these and many other people I owe a great debt of gratitude in helping to bring forth whatever is right in this book. Any errors of fact, logical lapses, omissions, uninterpretable utterances, and incomprehensible gaffes are, of course, entirely of my own doing.

This page intentionally left blank

Contents

This page intentionally left blank

Beyond Reduction

This page intentionally left blank

Introduction

Philosophical writing speaks in a number of different voices. Often, when we think of "philosophy," we think of works that present grand and original philosophical views about the nature of reality, knowledge, or morality. Familiar works cast in this mold include such notable examples as Kant's Critiques, Hegel's *Phenomenology of Spirit*, and Aquinas's *Summa Theologiae*. Other philosophical works take the form of commentaries upon, or critiques of, other intellectual discourses. These usually carry the label "philosophy of X," where 'X' denotes the other intellectual project with which the philosopher is critically engaged. In philosophy of science, for example, the sciences themselves are really the primary disciplines, and the contributions of philosophers consist either in the interpretation of the implications of these or else in criticism of their blind spots and shortcomings. A third type of philosophical writing attempts to work a kind of intellectual diagnosis and therapy upon ideas that are themselves philosophical, often the prevailing philosophical positions of the day. Much of Plato's corpus, especially the early, "Socratic" dialogues, is cast in this mold, as are many of the writings of the ordinary language philosophers and Wittgenstein.

This book has elements of each of these types of philosophical writing. Its primary intent is to work a certain amount of philosophical therapy upon a set of currently influential ideas in the philosophy of mind: namely, the suppositions that the mind must be "naturalized," and that the way to do this is to *reduce* mental states and processes to something else—something that can be captured in the language of physics, neuroscience, or other natural sciences. The contemporary debate between reductive naturalists and their principal opponents (nonreductive materialists, eliminativists, and dualists)

tends to proceed on the assumption that intertheoretic reductions are the norm in the sciences: that chemistry is reducible to physics, biology to chemistry and physics, and so on. Against this backdrop, the mind stands out as a striking anomaly. The centrally important properties of the mind, such as consciousness, intentionality, and normativity, do not seem to be reducible to what the brain does, or indeed to any facts specifiable in the languages of the natural sciences.

This problem, variously known as the "explanatory gap" (Levine 1983) or the "hard problem of consciousness" (Chalmers 1996), was posed almost four centuries ago by Descartes and has regained a good deal of notoriety in recent years. Is the appearance that there is such an explanatory gap merely a symptom of the current immature state of the sciences of the mind? Or perhaps of philosophers' ignorance of recent work in those sciences? Or is it a real and abiding feature of our understanding of the relationship between ourselves and the world of nature? And if so, if there is a principled limit to our ability to *understand* and *explain* the mind in terms of something else, what does this entail? Does it imply some form of dualism? Or that our ways of conceiving of the mind are so misguided that they do not in fact really refer to any real phenomena at all? Or perhaps merely that there are limitations to our own understanding that prevent us from having the same kind of insight into the basis of our own thinking that we have into things like atoms and metabolic processes? If we assume that intertheoretic reduction is the rule in the sciences generally, the explanatory gap would seem to be a crucial philosophical linchpin upon which our understanding of our place in the universe turns.

My contention in this book is that this entire problematic is misguided and is an artifact of an erroneous view in the philosophy of science. The crucial error is to assume that intertheoretic reductions *are* in fact the norm in the sciences (an error that was shared by proponents of the reductionist orthodoxy in philosophy of mind and by its challengers responsible for the resurgence of interest in the explanatory gap). This view was, to be sure, a central philosophical orthodoxy in the middle parts of the twentieth century. Yet over the past several decades, it has been decisively rejected within philosophy of science itself, and for reasons having nothing to do with the special problems encountered in examining the mind and its relationship to the brain. Biology is *not* reducible to chemistry and physics in the fashion conceived by such twentieth-century luminaries as Rudolf Carnap and Ernest Nagel. Indeed, in the relevant sense of 'reduction', chemistry is *not* reducible to physics, and thermodynamics is *not* reducible to statistical mechanics. Philosophy of mind at the turn of the millennium is, as it were, one of the last bastions of 1950s philosophy of science, and all parties to mainline debates about the nature of the mind err in making the assumptions (a) that the mind is unique in its irreducibility, and (b) that explanatory gaps are found only with respect to mental phenomena like consciousness and intentionality. There may, indeed, be special problems about the mind that are not encountered elsewhere; but irreducibility is not among them. The mind *is* irreducible; but it is hardly unique in this regard. Indeed, in some sense, in the sciences it is *explanatory gaps all the way down*.

If this is so, then a great deal in contemporary philosophy of mind is in need of some very fundamental rethinking. It is not merely the familiar positions, like dualism and reductionism, or the arguments for and against them, that need to be rethought. Rather, it is *the entire problematic* that is premised upon the assumption that intertheoretic reductions are widespread and that the mind is unique in resisting such reductions. The main point of this book is to drive home the point that post-reductionist philosophy of science ought to occasion some serious rethinking in philosophy of mind. In the course of this, some of the mainline contemporary problems in philosophy of mind are dissolved. They are, of course, replaced by new problems, such as *why* the philosophical project of reductionism failed, and how we ought to reconceive the mind and its relation to the world of nature.

This book does not attempt a definitive resolution of these problems, but it does present the lineaments of an alternative approach, called Cognitive Pluralism. Cognitive Pluralism is first presented as a thesis in philosophy of science, as an answer to the question of why the sciences are "disunified" in the sense of not being reducible to basic physics. My suggestion is that this is best understood by considering the sciences as *cognitive* enterprises: enterprises of modeling local features of the world (and of ourselves) in particular representational systems. Such models are local and piecemeal. They are also *idealized* in a variety of ways that can present principled barriers to their wholesale integration into something like a single axiomatic system. This view might be seen, in one respect, as a kind of generalization of the "Mysterian" views offered by Colin McGinn and Stephen Pinker. Whereas McGinn and Pinker suggest that the *psychological* explanatory gaps might be a consequence of limitations of our cognitive faculties, I suggest that the gaps represented by failures of reducibility in the science of natures might be understood in much the same fashion.

All of this might sound like a perfectly sensible move in the familiar direction of nonreductive physicalism: everything might *supervene* upon basic physics, and yet our minds may prove incapable of a global *understanding* of these supervenience relations in the form of the kind of axiomatic reconstruction of the special sciences envisioned by Carnap or Nagel. However, I think this would be the wrong conclusion to draw, for a number of reasons. First, I contend that, once one has rejected reductions, one no longer has a basis for preferring physicalism to its alternatives either. Second, the cognitivist turn involved in Cognitive Pluralism has metaphysical implications of its own. We can no longer rest content with a naïve realism that assumes that the world divides itself in a unique, canonical, and mind-independent way into objects and properties. Rather, the ways we carve up the world are inextricably bound up with the ways minds like ours represent features of the world. While Cognitive Pluralism may leave the *inventory* of the world as conceived by the sciences or by common sense essentially untouched, it cannot take that inventory as ontological bedrock. Like Kantian Idealism and Pragmatism, it embraces a "critical ontology" that asks *what it is to be an object*, and gives a cognitivist answer to the question. Third, both the cognitivist and the pluralist

strands of Cognitive Pluralism turn out to raise some fundamental issues for the practice of necessitarian metaphysics, especially as interpreted through possible-worlds semantics, and raise suspicions about our intuitions concerning things like metaphysical necessity and supervenience.

This book also undertakes a limited amount of philosophy of science. Most of this is framed at the level of exposition of several currents of existing work in philosophy of science that have helped to overthrow the reductionist orthodoxy of the 1950s. In the interest of moving along the main argument, I have opted not to argue these points anew, but instead to treat them as established conclusions in philosophy of science. It is, of course, possible that the next generation will show that the antireductionist trend witnessed over the past few decades in philosophy of science has been a mistake. The reader who suspects, hopes, or fears that this might be the case is invited to explore the books and articles referenced in chapter 3 and draw his or her own conclusions. I am content in this book to explore the rhetorical line that, *if* post-reductionist philosophy of science has it right, *then* philosophers of mind need to do some fundamental rethinking.

My aspiration for this book is threefold. First, I hope to bring philosophy of mind into closer dialogue with contemporary philosophy of science. I think that, for a good number of philosophers of mind, this *aim* will prove congenial, even if the conclusions prove surprising or even alarming. Second, I hope to introduce Cognitive Pluralism as an attractive approach in both philosophy of science and philosophy of mind. This book has not undertaken a full-scale exploration of Cognitive Pluralism. That will have to wait for another occasion. However, the basic lineaments of the position may prove sufficiently well-developed here for it to be deemed to merit further exploration. Third, and most fundamentally, I hope to provide comfort and solace for those, both in the profession and in the educated public, who think that reductionism is somehow implied either by the current state of the sciences or by the best philosophy of science available. I think that this assumption is widespread, but false. Indeed, I regard reductionism as a doctrine both false and harmful. I am not sure what I would do if I thought it harmful but *true*. Happily, I am not in that position.

Overview of the Book

This book is divided into three parts. Part I sets out some background on the problems and frames the terms of debate, hopefully in a way that will provide both a useful systematization for fellow specialists and an accessible point of entry for nonspecialists. Chapter 1 examines a variety of views that go by the name of "naturalism" in philosophy of mind, and contrasts them with the use of the word 'naturalism' in other areas, such as epistemology and philosophy of science. I argue that naturalistic philosophy of mind involves two kinds of claims: that mental phenomena can be *explained* in naturalistic terms, and that mental phenomena are *metaphysically supervenient upon and determined by* the

phenomena encountered in the natural sciences. I also argue that there is good reason for the fact that specifically *reductive* forms of naturalism have enjoyed pride of place in philosophical discussions, on the grounds that what I call "broadly reductive explanation," and *only* that form of explanation, guarantees metaphysical supervenience as well.

Chapter 2 undertakes a survey of the principal positions on the current scene in philosophy of mind: reductive and nonreductive materialism, eliminativism, dualism, and Mysterianism. These are presented in terms of the answers they give to four questions:

1. Can the phenomena of the (nonmental) special sciences be reductively explained?
2. Do the phenomena of the (nonmental) special sciences supervene upon the physical facts?
3. Can all mental phenomena be reductively explained?
4. Do all mental phenomena supervene upon the physical facts?

All parties involved tend to answer *yes* to the first two questions, and it is against this background that the explanatory gaps we seem to find with respect to the mind appear to present special and fascinating problems. Chapter 3, however, argues that the reductionist assumption reflected in a positive answer to the first question is in fact a kind of holdover from an outdated orthodoxy in philosophy of science. That chapter presents an overview of movements in philosophy of science that have resulted in the widespread rejection of intertheoretic reduction as a metatheoretical norm, and even of the assumption that such reductions are widespread in the natural sciences.

Part II then addresses the implications of post-reductionist philosophy of science for philosophy of mind by examining each of the familiar positions in turn. Chapter 4 examines reductionism and eliminativism, which are clearly compromised by any abiding pluralism in the natural sciences. If intertheoretic reductions are rare even in the natural sciences, there is little reason to expect them in the case of the mind, nor to hold the sciences of the mind in special suspicion because of their irreducibility. Chapter 5 examines the prospects of dualism, and chapter 6 those of nonreductive materialism and Mysterianism. I argue that each of these positions faces substantial problems in the wake of post-reductionist philosophy of science. On the one hand, their acceptance of the explanatory gaps in psychology is made more plausible by the realization that such gaps are indeed commonplace. But on the other hand, their evidential status is thrown into question, and along with it their ability to compete successfully either with other traditional positions or more radically pluralistic views that seem to be suggested by scientific pluralism.

Part III turns to the possibility that an abiding *explanatory* pluralism may point to a need to explore a more systematic *philosophical* pluralism. Chapter 7 discusses two types of pluralism. The first is Dupré's (1993) "promiscuous pluralism," a kind of realist pluralism with a radically expanded ontological inventory. The second is the view I wish to recommend, Cognitive Pluralism.

Cognitive Pluralism is first discussed in chapter 7 in epistemological terms, as a possible explanation of why there might be abiding explanatory pluralism in the sciences. Chapter 8 then argues that the key notions developed in chapter 7 —especially that the mind understands the world through special-purpose, idealized models—is not a feature distinctive of *scientific* understanding so much as it is a general feature of human cognitive architecture, of which scientific understanding is but a particularly exacting and regimented case. In both of these chapters, it is argued that the use of special-purpose, idealized models, each employing a representational system suited to its individual problem domain, (a) may be a deep "design principle" of human cognitive architecture that cannot be avoided, and (b) is sufficient to explain some types of abiding disunities in our knowledge as artifacts of our cognitive architecture. Cognitive Pluralism is then discussed as a metaphysical thesis in chapter 9. There it is argued that both its cognitivist and its pluralist strands give us reason to rethink the status of intuitions about claims for metaphysical necessity and supervenience that have shaped recent discussions in the metaphysics of mind. Chapter 10 returns to the topic of naturalism, and asks whether a naturalist might also be a Cognitive Pluralist, and vice versa. The answer to this depends upon the operative sense of the word 'naturalism'. If it is used as it is employed in philosophy of science and epistemology—that is, as signifying a rejection of aprioristic theories in favor of theories more engaged with the sciences themselves—then Cognitive Pluralism is intended as a paradigm example of a "naturalistic" approach. But if it signifies the view that there is a single privileged set of "natural" facts upon which all of the others depend, and from which they may be derived, Cognitive Pluralism is a radical repudiation of naturalism.

PART I

Naturalism and Reduction in Philosophy of Mind and Philosophy of Science

This page intentionally left blank

I

Varieties of Naturalism

What Is a Naturalistic Philosophy of Mind?

A casual observer of recent philosophy of mind would likely come to the conclusion that, amid all of the disagreements between the parties involved, there is at least one thing that stands as more or less a consensus view: the commitment to a philosophy of mind that is *naturalistic*. Almost everyone writing in philosophy of mind over the past several decades has described his or her theory as ''naturalistic.'' This includes proponents of quite a wide variety of views: reductionist, informational, nonreductive physicalist, functionalist, and evolutionary. Even David Chalmers, perhaps the most influential figure in the revival of property dualism in the late 1990s, describes his position as ''naturalistic'' (Chalmers 1996). At first glance, then, philosophers of mind might seem to have found at least one happy point of agreement at the turn of the millennium.

But things are not so simple. And the fact that they are not so simple ought to be foreshadowed by the very variety of views that can be styled ''naturalistic.'' If a reductionist, an evolutionary theorist, and a dualist can each apply the label 'naturalist' to himself or herself, it is very likely to prove the case either that they are using the word in subtly different ways, or else that the word has become so bland and ecumenical as to be essentially useless.

I am not really pointing out anything new here. The ambiguity of the word 'naturalism' has been widely noted, and has been remarked upon for perhaps half a century now. The midcentury philosopher of science Ernest Nagel, in his 1955 presidential address to the American Philosophical Association, noted that ''the number of distinguishable doctrines for which the word 'naturalism' has been a counter in the history of thought is notorious'' (3). In their introduction to the

anthology *Naturalism: A Critical Appraisal*, Wagner and Warner (1993, 3) express a similar view:

> Participants in current discussions of naturalism seem to assume that the meaning of 'naturalism' ('naturalist program', etc.), its motivations and—often—its correctness, one way or the other, are almost obvious. The historical situation makes such assumptions exceedingly unlikely. Philosophers have taken just about every possible stance with some manner of justification, and all of the main programs within this area ("naturalism," "phenomenology," "analytic philosophy," and so forth) have been open to sharp differences of interpretation by their adherents.

In a similar vein, David Papineau (1993, 1) begins his book *Philosophical Naturalism* with the words, "What is philosophical 'naturalism'? The term is a familiar one nowadays, but there is little consensus on its meaning. . . . I suspect that the main reason for the terminological unclarity is that nearly everybody nowadays wants to be a 'naturalist', but the aspirants to the term nevertheless disagree widely on substantial questions of philosophical doctrine." Some philosophers, such as Jesse Hobbs (1993), have taken Papineau's point that "nearly everybody wants to be a 'naturalist' " even further, raising the question of whether the word 'naturalism' is simply "a contemporary shibboleth." If one came to this conclusion, one would, I think, be *half* right. The word 'naturalism' *does* tend to function as a kind of shibboleth—that is, as a word whose use distinguishes "members of the tribe" from outsiders. And it is, I think, true that naturalism has become a kind of ideology in philosophical circles; that is, it is a widely shared commitment to a way of believing, speaking, and acting whose basic assumptions are seldom examined or argued for. However, I think that this is not the whole story. The word 'naturalism' may serve as a shibboleth, but it is not *merely* a shibboleth. There may be a pervasive naturalistic ideology that masks a variety of more specific views, but it is possible to articulate and examine some basic shared underpinnings. And if there is not a *single* view called "naturalism" shared by the majority of contemporary philosophers of mind, there is nevertheless a way of bringing some order to the various views thus described, highlighting their commonalities as well as their differences.

1.1. 'Naturalism': A First (Inductive) Characterization

Like most philosophers, I try to discourage my students from using nonphilosophical dictionaries as authorities on the meanings of philosophical terms. In the present case, even the best of English dictionaries, the *Oxford English Dictionary*, misses the subtleties of philosophical usage. However, it does, at the very least, help us in distinguishing philosophical usages from various nonphilosophical usages of 'naturalism' and 'naturalist', such as "An expert in or student of natural science."[1] The OED's entry on the (nonethical) philosophical usage is also useful, not least for its historical material:

> **2.** *Philos.* The idea or belief that only natural (as opposed to supernatural or spiritual) laws and forces operate in the world; (occas.) the idea or belief that nothing exists beyond the natural world. Also: the idea that moral concepts can be analysed in terms of concepts applicable to natural phenomena. Cf. NATURALIST *n.* 2a.
>
> **1750** W. WARBURTON *Julian* 42 *note*, [Ammianus] being. .a religious Theist, and untainted with the Naturalism of Tacitus. **1794** R. HURD *Life Warburton* 72 Lord Bolingbroke. .was of that sect, which, to avoid a more odious name, chuses to distinguish itself by that of Naturalism. **1816** R. HALL *Let.* in *Wks.* (1832) V. 502 Their system is naturalism, not the evangelical system. **1858** E. H. SEARS *Athanasia* 4 By the word 'Naturalism' we describe a belief in nature alone. **1874** W. WALLACE *Logic of Hegel* §60. 100 Materialism or Naturalism, therefore, is the only consistent and thorough-going system of Empiricism. **1903** G. E. MOORE *Principia Ethica* ii. 40, I have thus appropriated the name Naturalism to a particular method of approaching Ethics. **1967** *Encycl. Philos.* III. 69/1 According to ethical naturalism, moral judgments just state a special subclass of facts about the natural world. **1972** N. MCINNES *Western Marxists* i. 25 Marxism begins as pure philosophy but it has a tendency to 'degenerate' into social naturalism. **1992** *Mind* **101** 131 Armstrong advocates Naturalism: 'the doctrine that nothing at all exists except the single world of space and time'.

This definition and the attendant quotations tease out several themes that do indeed seem to play a large role in the forms of naturalism found in philosophy of mind. (It is less clear that this definition fits the usage of 'naturalism' in epistemology or philosophy of science.) The first of these is a *metaphysical* claim to the effect that the inventory of the *natural* world exhausts the inventory—or at least the *basic* inventory—of the world *simpliciter*. The second is more of a claim about *epistemology*, analysis, or explanation: that things that do not, on the face of it, seem to be parts of nature (such as minds and norms) can in fact be understood in terms of natural phenomena. Implicit in all this, moreover, are two additional assumptions. The first is that "the natural" is to be understood as *the domain(s) of the natural sciences, particularly physics*. The second is that there is an implicit contrast class for "naturalistic" theories: they are identified by contrast with theories that appeal to "supernatural" or "spiritual" laws or forces.

We may thus make a first attempt at a schema for at least one influential philosophical notion of "naturalism."

> *Naturalism—a General Schema*: Naturalism about domain D is the view that all features of D are to be accommodated within the framework of nature as it is understood by the natural sciences.

Thus naturalism in philosophy of mind would be (on first approximation) the view that all mental phenomena are to be accommodated within the

framework of nature as understood by the natural sciences, naturalism in ethics would be the view that all ethical facts are to be accommodated within such a framework, and so on for any domain to which this schema is applicable.

1.2. Three Dimensions of Ambiguity

This view is not really anything so exact as a shared *theory*. Instead, it is something on the order of a theory-*schema*. It is only a schema for theories because there are several elements of this characterization that are ambiguous, and which different self-styled "naturalizers" would fill out in different ways. In addition to the question of what domain is to be naturalized (e.g., mind or ethics), there are (at least) three axes along which this schema is ambiguous that can be used to differentiate varieties of naturalism:

1. Whether the "accommodation" in question is a sort of *explanation* or a type of *metaphysical determination*
2. How we are to understand "the framework of nature as it is understood by the natural sciences," and
3. Whether the general schema is understood as a *positive* claim (that the mind *can* be so accommodated) or as a *normative* claim (that it *must* be so accommodated, or else some dire consequences follow).

Let us consider these issues in order. Examinations of naturalism in philosophy of mind often mix together discussions of whether features of the mind such as consciousness and meaning can be *explained* by the natural sciences with discussions of *metaphysical* questions (such as whether mental states supervene upon brain states). For many naturalists, both sorts of questions are deemed to be of great importance. And there *are* styles of explanation that are closely linked to particular types of metaphysical determination. However, metaphysical questions and questions about explanation are separable from one another. On the one hand, there are forms of explanation, such as statistical explanation, that have no metaphysical consequences. On the other hand, it might be the case that there are metaphysical necessities that are *epistemically opaque*—that is, they are necessarily true, but in such a fashion that our minds cannot understand *why* they *must* be so—and which consequently have no attendant forms of explanation to go along with them that underwrite their necessitarian character. (Many nonreductive physicalists and Mysterians, for example, believe that mental phenomena *supervene* upon facts about the brain, but cannot be *reductively explained* by them.) So in examining a particular naturalistic claim it will be important to identify whether it is a claim about explanation, or a claim about metaphysics, or both.

Likewise, even once we have pinned down what we mean by "accommodating" the mind within nature, the expression "the framework of nature as it is understood by the natural sciences" is still rather vague. Just what our

naturalistic schema *means* will depend heavily upon what one considers to be central to how the natural sciences operate, and how they represent the natural world. That is, it will depend upon what particular views one takes in philosophy of science on issues like the nature of explanation and the metaphysical commitments of the sciences. And this is a serious complication, because there are many alternative views on these subjects, as we shall see in section 1.3.

There is also a third axis of ambiguity: sometimes naturalistic claims are put forward as a kind of *positive* claim—a claim about how things are. These are a sort of second-order empirical claim about how it will turn out in the long run. Positive empirical claims can often be put to the test and be shown to be true or false: it might turn out that some feature of the mind, such as consciousness, *can* be naturalized, or it might turn out that it *cannot*. But some naturalists have an uneasy tendency to slide into a different sort of claim that is *not* empirical or positive, but *normative*. They claim, in essence, that the mind *must* be naturalized, or else something unseemly follows: that psychology cannot be scientific unless its objects can be explained in terms of something more fundamental, or that mental states do not exist unless they supervene upon physical states. Stephen Stich and Stephen Laurence (1994, 160) describe the situation in the following way with respect to the particular project of naturalizing intentionality:

> In recent years, many philosophers have put a very high priority on providing a "naturalistic" account of intentional categories. Moreover, there is an unmistakable tone of urgency in much of this literature. Naturalizing the intentional isn't just an interesting project, it is vitally important. *Something dreadful* will follow if it doesn't succeed. And for many writers, we suspect, that dreadful consequence is intentional irrealism.

Positive and normative claims must be evaluated in very different ways, and so it behooves us to be careful in identifying which sort of claim we are dealing with.

Additionally, our schematic characterization requires an important caveat. Depending on what view one takes of what it would mean to "accommodate" some phenomenon within "the framework of nature as it is understood by the natural sciences," this formula might let in views that would be paradigmatically *non*naturalistic. For example, if one were to follow Jaegwon Kim (1993) in equating "the natural" with things that enter into causal relationships, this would include as "natural" objects both a God who created the universe and Cartesian immaterial souls that entered into causal relations with human bodies. But construing naturalism *this* broadly would leave no meaningful contrast between naturalistic and nonnaturalistic views. Thus we should augment our general schema with the following caveat:

> **Caveat:** a naturalistic theory cannot be one that
> (a) posits the existence of supernatural entities (such as God, angels or immaterial souls), or

(b) adopts a metaphysical stance in which the ontology of the natural sciences is not fundamental (e.g., transcendental idealism, pragmatism).[2]

1.3. "The World of Nature as Understood by the Natural Sciences": Three Views

Two of the ways our schema is ambiguous require little additional comment at this point. It is clear enough what it means to say that questions about metaphysics need at least initially to be distinguished from questions about explanatory success, though of course the relationship between certain types of explanation (particularly reductive explanation) and metaphysics will need to be taken up at a later point. Likewise, it is clear enough what it means to distinguish claims made in the assertoric voice, as second-order empirical hypotheses about how the sciences of the mind can be united with the natural sciences, from those made in the normative voice and intended to serve as a kind of constraint upon psychology or philosophy of mind.

By contrast, it is necessary to say a bit at the outset about different views of what might be understood by "the world of nature as understood by the natural sciences." Some would-be "naturalizers" of the mind are *reductionists*. Others are concerned with *lawlike* relations between mind and body, or among mental states. And still others wish to understand the mind in biological terms, employing resources from *evolutionary theory* or sociobiology. These three approaches really reflect three different views of scientific explanation, which may be associated in turn with three pivotal figures in the history of science: Galileo, Newton, and Darwin.

1.3.1. Galileo and Reduction

Galileo made important contributions, not only to mechanics, but also to scientific methodology. His approach is often called the Method of Resolution and Composition (MRC). The basic idea of the MRC is that, to understand a complex phenomenon, one first must break it down into its component parts. (This is the *resolutive* or *analytic* step.) Then one examines the properties of the parts and tries to derive the observed behavior of the larger system from the assumptions about the parts. (This is the *compositive* or *synthetic* step.) Explanation is completed when it is possible to derive all relevant features of the system you are trying to explain from properties of the parts. This method, also endorsed by Galileo's younger contemporary Descartes and by his sometime visitor Thomas Hobbes, became a mainstream tenet of seventeenth- and early eighteenth-century mechanism, and was revived in a slightly different form in the twentieth century by the Logical Positivists and Empiricists, who called the view 'reductionism', a shorter if less informative label than Method of Resolution and Composition. From there it played an important role in the development of both analytic behaviorism and reductive physicalism in philosophy of mind.

In both its Early Modern and its Positivist forms, reductionism took its inspiration from the methods of mathematics. The Early Moderns used geometry (analytic geometry, in the case of Descartes) as their paradigm, and viewed explanation on the model of mathematical deduction and/or construction. The Positivists preferred the model of the logical syllogism. Reduction was supposed to be a very complete and rigorous form of explanation, comparable to mathematical demonstration. In its strongest forms, a reduction of, say, chemistry to physics or of thermodynamics to statistical mechanics was supposed to resemble a mathematical proof or a derivation in an axiomatic system.[3] Given the assumptions about the more basic system, the salient features of the reduced system could be inferred completely and with confidence. As a result, reduction ensures a strong metaphysical relationship as well. If one can deduce or construct everything about a complex system A from assumptions about its parts B, then A is metaphysically supervenient upon B as well, and B→A is metaphysically necessary. (The notion of supervenience will be discussed further in chapter 2.)

1.3.2. Newton and Laws

In the eighteenth century, followers of Isaac Newton tended to reject the reductionistic model of science. In an oft-cited passage in the General Scholia to the second edition of the *Principia*, Newton (1713/1962, 546–47) wrote: "Hitherto we have explained the phenomena of the heavens and of our sea by the power of gravity, but have not yet assigned the cause of this power. . . . But hitherto I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypotheses." Exactly what Newton himself understood by "I frame [or, in a better translation "feign"] no hypotheses" ("hypotheses non fingo") is a matter of lively scholarly debate. However, Newtonians like Hume tended to take this as licensing a rejection of the search for hypothetical unseen mechanisms in favor of a search for mathematical laws that describe the observable phenomena. Science was not to be in the business of postulating mechanisms so much as finding laws that would allow for prediction and control. Newton's own view as expressed in the *Scholia* seems to have been that we need to postulate a gravitational *force* but that he had no hypotheses about the *mechanism* by which such a force might operate. His philosophical interpreters tended to draw the more radical moral that we ought not to postulate *forces*, but merely to take laws as systematizing the phenomena, placing Newton's laws more on a par with Kepler's. (This is one of several respects in which Newton himself was, in my view, a better philosopher of science than Locke or Hume. Compare Schliesser 2004.)

As a result, the eighteenth- and nineteenth-century attempts to extend Newtonian methodology to psychology, from Locke to James Mill, tended to eschew the search for mind-body connections in favor of a "mental chemistry" or "mental geography" that would uncover laws (generally understood to be laws of association) linking one mental state to another. Such a law-centered vision of science was in the later nineteenth century extended to the connections

between percepts and the stimuli that cause them by psychophysicists like Weber and Fechner, and also in crossover work by the physicists Mach and Helmholtz. Several contemporary philosophers have tended to speak of the relationships between mental states and their corresponding brain states as "psycho-physical laws" as well, albeit in a different sense from the "psycho-physical laws" discovered by Weber (e.g., Davidson 1970; Chalmers 1996).[4]

If "naturalizing" the mind consists in finding *laws* relating (a) pairs of mental states to one another, and/or (b) mental states to stimuli, and/or (c) mental states to brain states, and/or (d) mental states to behavior, the naturalist has an agenda that falls considerably short of Galilean-style reduction. Reductive connections, modeled on mathematical demonstration and construction, carry the force of metaphysical necessity. By contrast, laws, even physical laws, are generally held to be metaphysically contingent. They consist first and foremost in robust empirical generalizations about things that co-occur, and often involve the postulation of causal connections between the events related by the law. (When this is so, it is not a reduction or an identity, as causation is a relation between two distinct events rather than one event under two descriptions.) Finding an empirical generalization that relates A and B does not preclude a reduction of A to B, or the identification of A with B, but it does not entail them either, and indeed if such a reduction or identity relation were to be found, we might well cease to speak of the relation as a law.

A merely nomic connection between mind and body, moreover, is compatible with a variety of metaphysical interpretations. It is compatible with physicalism. But it is also compatible with property and substance dualism, interactionism, and for that matter with various forms of idealism, pragmatism, neutral monism, and social constructionism as well. It thus would seem to violate the Caveat offered earlier that disqualifies views that countenance things like Cartesian souls from being labeled "naturalistic." One can, of course, choose to use the word 'naturalism' in a weaker sense, as Chalmers (1996) does, for example. In part, this is merely a dispute over words; but this usage seems to go against the spirit and motivations historically associated with the use of the word, both historically and on the contemporary scene. And so I shall take the view that *a merely nomic form of "naturalism" is really no naturalism at all*, especially as it would have been endorsed by someone like Descartes, so often identified as a principal opponent of naturalism, who nevertheless thought there were nomic *causal* relations between mind and body. Nomic relations are *compatible* with the Caveat, and hence with naturalism, but they are not themselves sufficient to ground naturalism.

1.3.3. Darwin and Evolutionary Explanation

Still other proponents of views styled "naturalistic" are interested in accommodating the mental under the aegis of evolutionary biology. In its mildest form, Darwinian naturalism treats specific types of mental states—pains, desires, beliefs—as phenotypic features of an organism that are to be explained through mechanisms of variation and selection at work in the ancestral history

of the species. There are four elements to such an evolutionary story: (1) how the trait initially comes on the scene through some process of spontaneous variation, (2) how it is heritable from one generation to another, (3) how it is expressed in an organism through development, and (4) how mechanisms of selection account for its proliferation as an adaptation. Evolutionary psychology and sociobiology generally concentrate on the final element, telling stories about the hypothesized adaptive value of various mental traits. Some forms of Darwinian naturalism go further than this. Millikan (1984), for example, attempts to account for the *nature* of mental traits through biological explanation. The nature of a trait is its proper function. Likewise, Dretske (1995) differentiates what a mechanism in an organism *actually* does from its *function*, understood as what it was *selected* to do.

Evolutionary explanations of the mind are sometimes viewed as closing an important gap between physics and psychology, and thus as providing a necessary supplement to reductive explanation. However, this is misleading. Stories about the adaptive value of a phenotypic trait bring that trait within the broader scope of the natural world only when supplemented with the rest of the evolutionary story about the appearance, inheritance, and expression of the trait (Horst 1999). Consider two extreme examples. An organism that was supplemented with a Cartesian rational soul would likely enjoy competitive advantages over organisms lacking such a soul if such a soul conferred the benefits Descartes suggested: rationality and language (cf. *Discourse V* in Descartes 1985, vol. 1). You could tell a good *adaptational* story about Cartesian souls. However, having an immaterial rational soul is not the sort of thing that could be transmitted genetically to one's offspring, and hence is not a trait on which gene selection could operate. Likewise, it is not the sort of thing that could be the result of the expression of genes. Or consider a second example: an organism that was powered by a perpetual-motion machine would have an enviable degree of differential fitness in that it would not need to eat to live, and hence would be immune to famine and could devote more of its energies to producing offspring. However, we have good reason to suppose that a physical world like our own could not endow an organism with a perpetual-motion machine, and hence could not supply the preconditions for forces of selection to operate.

The moral of these (admittedly extreme) stories is that evolutionary explanations are suspect in precisely the cases where there is reason to wonder whether merely physical mechanisms *could* indeed produce the phenotypic trait in question. Settling the question of whether physical mechanisms can do so in a given case is precisely what is at stake in reductive explanations. To the extent that one has reason to doubt that a mental trait is indeed subject to reduction, one thereby has reason to doubt that it is something that could arise through mutation, be expressed through development, or upon which mechanisms of selection could operate. They are *defeasible* and *prima facie* reasons to doubt it, as it might be possible to provide nonreductive explanations of mutation, heritability, and development. But more generally, arguments to the effect that mental phenomena like consciousness and meaning

cannot be accounted for by the physical phenomena going on in the brain are by extension arguments against evolutionary accounts of the mind as well. Concentrating on the *selectional* component of evolutionary explanation creates the illusion of bypassing the problems of alleged explanatory and metaphysical gaps; but an illusion it is. We cannot tell a Darwinian story about the inheritance or selection of a trait unless it is something that could be the result of the expression of genes in development and passed on through physical mechanisms of inheritance, and these are precisely what such antinaturalistic arguments call into question.

1.4. The Preeminent Importance of Reductive Naturalism

A merely nomic form of naturalism—one that treats "naturalization" of the mind merely as a task of finding laws that relate mind and brain—is really no naturalism at all. The viability of Darwinian naturalism is ultimately dependent upon exactly the questions that are at stake in reductive naturalism. As a result, this book concentrates on reductive forms of naturalism. On the one hand, it is they, and only they, that license a move from successful explanation to a metaphysical conclusion such as physicalism or metaphysical supervenience. On the other hand, nomic and evolutionary naturalism do not, by themselves, really bring the mental within the sphere of the physical world. In the case of nomic naturalism, its claims are compatible with alternative metaphysical construals, such as dualism and idealism. In the case of evolutionary naturalism, its plausibility is ultimately dependent upon the very issues that are at stake between reductive naturalists and antinaturalists. It is therefore reduction that is the focus of our investigation here.

1.5. Other Philosophical "Naturalisms"

While this chapter is not intended as a general encyclopedia entry on 'naturalism', a few words on uses of the term in philosophical specialties other than philosophy of mind are in order. The word has played an important role in ethics for roughly a century now, in epistemology for about half that time, and in philosophy of science for several decades. The usage in ethics has important parallels with that in philosophy of mind, but those in epistemology and philosophy of science are importantly distinct from, and in some ways in tension with, the usage we have already explored.

The term 'naturalism' was introduced into ethics by G. E. Moore (1903). The naturalism Moore opposed was largely constituted by attempts to "derive *ought* from *is*"—that is, to try to construct normative notions from purely positive notions. While I suppose that someone might try to do this within a dualist metaphysics—say, to treat norms as a consequence of positive facts about immaterial souls—for the most part such attempts are cashed out in terms continuous with the natural sciences. And ethical naturalisms tend to

suffer at least one of the same ambiguities found in philosophy of mind. Sometimes the "naturalism" in question is a claim at the level of *explanation*: that normative claims can be *reduced to* nonnormative claims. But sometimes ethical "naturalism" signifies only the claim that normative facts supervene upon nonnormative facts, regardless of whether the ethical categories themselves can be reconstructed out of purely positive claims. Debates about ethical naturalism have a great deal in common with debates about naturalizing the mind. Some antinaturalists in philosophy of mind (e.g., Brandom 1994) hold that intentional states are constitutively normative, and hence arguments against the reducibility of the normative in ethics can be applied, *mutatis mutandis*, to intentional states as well.

The usage in epistemology dates to Quine's (1969) essay "Epistemology Naturalized." Quine's main concern was that epistemology should not be pursued as an armchair, aprioristic enterprise, but should be informed by, and indeed be continuous with, the scientific study of the mind. 'Naturalism,' here, primarily signifies a methodological position. Likewise, "naturalistic" philosophy of science is characterized by the attitude that the philosophy of science ought not to proceed by applying aprioristic standards (such as the Positivist conception of the "logical form of Science") that are then held as litmuses for good scientific practice. Instead, it should proceed by studying how (successful) work in the sciences in fact operates (cf. discussions in Callebaut 1993).

It is important to see that there is at least a potential conflict between the naturalistic approach to philosophy of science and at least one form of naturalism in philosophy of mind: namely, that which treats reducibility as a kind of norm used for testing the credentials of a special science like biology or psychology. To adopt such a normative stance is to indulge in exactly the kind of application of extrascientific standards that the naturalistic philosopher of science rejects. Normative naturalism in philosophy of mind and naturalistic philosophy of science make for poor bedfellows.

This is not simply an idle observation. The pivotal chapters of this book address the changes that have occurred in philosophy of science over the past several decades and their implications for the philosophy of mind. Much of the recent problematic in philosophy of mind has been shaped by assumptions arising from the Positivist and Logical Empiricist projects in philosophy of science, and in particular by the idea that intertheoretic reductions are the norm in the natural sciences and should perhaps be viewed as normative constraints upon an acceptable account of the mind. To the extent that such an assumption has been rejected within philosophy of science itself, this may give us reason to rethink the standard problems in philosophy of mind and to reevaluate all of the familiar philosophical accounts of the nature of the mind.

This page intentionally left blank

2

Reduction and Supervenience

The Contemporary Problematic in Philosophy of Mind

A great portion of the work done in philosophy of mind in recent years has been devoted to two broad questions. The first of these is an epistemological question: *Can mental phenomena such as intentionality and consciousness be reductively explained in terms couched in the languages of the natural sciences?* The second is metaphysical: *Do mental phenomena supervene upon the sorts of facts described by natural sciences such as physics, biology and neuroscience?*

2.1. Reductionism

The mid-twentieth century was a heyday of reductionism in philosophy of mind. Reductionists hold that mental phenomena supervene on physical facts, and are also reducible to physical facts. Such a position is generally seen as being *naturalistic* in its bent: that is, it is one form of the view that the mind can (or perhaps must) be accommodated within the framework of the world of nature as understood by the natural sciences. In the mid-twentieth century, the prevailing philosophical view was that the natural sciences themselves form a kind of natural hierarchy based on composition and complexity. The fundamental facts about the universe, as well as the fundamental laws, are those that are couched at the level of basic physics. The facts and laws of "special" sciences like chemistry and biology are necessitated by these, and moreover can be derived from them by way of intertheoretic reductions. To view mental phenomena as reducible and supervenient is simply to view the relationship between psychology and the natural sciences as

being analogous to the relations thought to obtain between sciences like chemistry or biology and basic physics.

Such a reductionist stance can be held in two very different sorts of ways. The Logical Positivists came to it by way of their interest in what they called "the logical form of science," an aprioristic standard that held the actual practice of the sciences up to philosophical norms about justification, explanation, and unity. Positivists viewed explanation within a science as a form of syllogistic argument from laws (interpreted as universally quantified claims ranging over objects and events) and initial states of objects, to their subsequent states. Relations between sciences were also viewed as being (or being reconstructable as) axiomatic systems in which all of the primitive definitions and axioms would be couched at the level of basic physics (or whatever a given Positivist took to be the fundamental science, as some of them were sense-data phenomenalists rather than physicalists), from which one could derive the truths of the special sciences in a fashion analogous to the derivation of constructions, lemmas, and theorems in mathematics. For many of the Positivists, it was something on the order of a rational norm that all scientific knowledge should be reconstructable in the form of such a grand axiomatic system. This closely reflected the attitude of Rationalistic mechanists like Descartes and Leibniz (and, in my opinion, Hobbes should be included in this camp as well) in the seventeenth and early eighteenth centuries. Indeed, Rationalists and British Empiricists alike still employed a Scholastic notion of *scientia* that more or less required the kind of demonstrative character found in geometry. (Even Hume was hesitant to call Newtonian laws "knowledge"—i.e., *scientia*—because they could not be known demonstratively.)

Other twentieth-century proponents of reductive unification viewed reductionism as a second-order empirical hypothesis rather than a rational norm. Most famously, Oppenheim and Putnam (1958, 7), in their "Unity of Science as a Working Hypothesis," speculated that such a system might plausibly be extended even to include psychology:

> It is not absurd to suppose that psychological laws may eventually be explained in terms of the behavior of individual neurons in the brain; that the behavior of individual cells—including neurons—may eventually be explained in terms of their biochemical constitution; and that the behavior of molecules—including the macro-molecules that make up living cells—may eventually be explained in terms of atomic physics. If this is achieved, then psychological laws will have, in *principle*, been reduced to laws of atomic physics.

There were several conjectures as to what form such a reduction of psychology might take. Carnap and several other Positivists at one time favored a view called "analytic behaviorism," "the doctrine that, just as numbers are (allegedly) logical constructions out of *sets*, so *mental events* are logical constructions out of actual and possible *behavior events*" (Putnam 1961/1980, 25). In spite of the advocacy of such distinguished philosophers as Carnap (and perhaps Ryle and

Wittgenstein), analytic behaviorism soon met insuperable obstacles, not the least of which was that in several decades, no one was able to produce so much as a single plausible analysis of any mental state in behavioristic terms.

By the 1960s, another reductionist proposal had begun to take hold: the (type) identity thesis of Place (1956) and Smart (1959). This thesis proposed that the reduction base for mental phenomena was not behavior events, but brain states. Its claim was that, for each type of mental state, there is a unique brain state with which it is identical. Whereas analytic behaviorists would have identified pains with their bodily causes (say, tissue damage) and behavioral effects (say, wincing, cries of agony, and withdrawal from the offending stimulus), type-identity theorists claimed that pains are identical with a particular type of neural event, conveniently labeled "C-fiber firings."

2.2. Functionalism

Type-identity theory was in its turn dethroned by the functionalist movement in the 1960s through 1980s. Functionalists pointed out that there are perfectly legitimate kind-terms whose characteristic features are not structural but functional. What is typical of hearts, *qua* hearts, is that they pump blood, or perhaps that they have the function of pumping blood. But not all hearts are physically alike. Earthworm hearts and human hearts bear few anatomical similarities at the structural or microcellular levels, but perform similar functions; artificial hearts perform the same function without being composed of cells at all. Particular functionally defined types of circuits, such as AND-gates employed in computers and other electronic devices, can be built out of a variety of types of materials, indeed from an infinite number of types. Functional types are thus said to be "multiply realizable": each functional kind stands in a one-to-many relation to the various sorts of physical systems through which the function may be realized (see Figure 2.1).

Functionalists argued that, just as human hearts are different from earthworm hearts, the physical systems that realize pains (or other mental state-types) in different species might be different as well. Indeed, if pain is characterized by functional rather than structural properties, one might plausibly suppose that there could be Martians or even robots that could have the selfsame *functional* states, but realized through things other than nerves or even cells at all (Lewis 1978). In spite of important objections (e.g., Block, 1978/1980), functionalism became the mainline view of the nature of mental states in the 1970s and 1980s, and type-identity theory was increasingly regarded as having been decisively refuted.

It is partly a matter of terminology whether functionalism should be viewed as an alternative to reductionism or a form thereof. There is a usage of the word 'reductionism' in philosophy of mind which means specifically the type-identity theory. However, the word has always had, and still enjoys, broader uses as well. Reductionism, broadly construed, really asks less than type-identity. Type-identity, in positing a one-to-one correspondence between

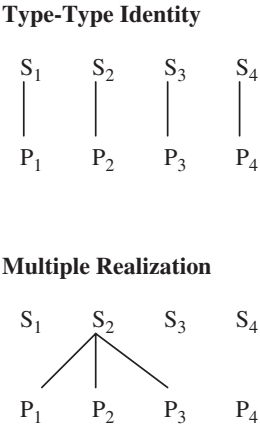


FIGURE 2.1. Type Identity and Multiple Realization. Type identity theorists held that each type in a special science, like biology or psychology, was identical with some distinct physical type. Functionalist proponents of multiple realization held that some types in the special sciences—those that are functional types—can be "realized" by more than one physical type. Type identity theory posits a one-to-one relation between psychological types and physical types, whereas multiple realization involves one-to-many relationships.

mental state types and physical state types, allowed both an *upward* inference from physical type to mental type and a *downward inference* from mental to physical type (see Figure 2.2). But the core of reductionism is really only the upward inference. And this is not only compatible with functionalism, but embraced by many of its advocates. Indeed, functionalists can be seen as pursuing a reductive programme at two levels. First, they postulate a type-identity between mental types and functional types. Second, they are generally of an opinion that an adequate understanding of a system at a physical level should, in principle, be sufficient to explain and to entail its functional properties. Functionalists are thus advocates of a "broad reductionism" whose nature will receive more careful explication shortly.

2.3. Eliminativism

In the 1980s, broadly reductionist views (then represented primarily by functionalism based on the computer metaphor) encountered opposition from a radical alternative. Eliminativists (e.g., P. M. Churchland 1981; Stich 1983; Ramsey, Stich, and Garon 1991) claimed that a more careful and scientific study of the mind through hard sciences like neuroscience would in fact turn up nothing corresponding to the "folk psychological" inventory of beliefs, desires, and qualia. They viewed all of these as "posits" of a "folk psychological theory" and suggested that they would ultimately go the way of other failed theoretical postulates, such as phlogiston, and be "eliminated" from our ontology once superseded by a more perspicuous typology arising from

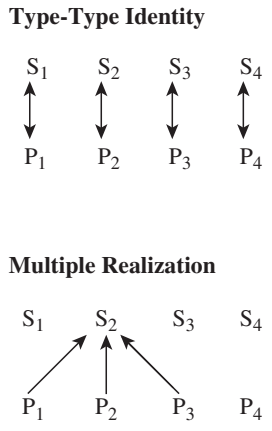


FIGURE 2.2. Inferences Allowed by Type Identity and Multiple Realization of Functional Kinds. Type identity allows both "upward" inferences from physical type to the types of psychology and other special sciences and "downward" inferences from special science type to physical type. Multiple realization allows only "upward" inferences, as the "downward" realization relation need not be unique.

neuroscience. Eliminativism in philosophy of mind is compatible with a broad reductionism about the natural sciences. Indeed, the prime reason for favoring an eliminativist view of the mentalistic vocabulary is the suspicion that such a vocabulary cannot be reduced to that of physics or even neuroscience. Eliminativists share with normative reductionists a commitment to the principle that the psychological phenomena *must* be reducible to facts of the natural sciences in order to be scientifically respectable and ontologically legitimate. But whereas psychological reductionists (including functionalists and computationalists) believe that such broad reductions are there to be had, eliminativists believe them to be chimerical.

2.4. The Explanatory Gap(s)

Throughout the 1980s, philosophy of mind was largely dominated by the assumption that there is a kind of forced choice between reduction and elimination. During this period, however, there were the beginnings of a backlash against this orthodoxy. Levine (1983) claimed that there is an "explanatory gap" in the form of an irreducibility of several important mental phenomena to physical objects and processes. And this view was supported by several compelling thought-experiments, such as Jackson's Mary,[1] Nagel's bat,[2] and Searle's Chinese Room.[3] Around the same time, Davidson (1970) argued on very different grounds that not only is the mental irreducible, but that there are not even any laws linking descriptions of the physical world to unique interpretations in terms of beliefs and desires. Such views were regarded primarily as interesting anomalies in the 1980s, but the explanatory gap gained increasing plausibility in the 1990s with

arguments for the irreducibility of consciousness and qualia (Chalmers 1996), intentionality (Searle 1992; Horst 1996; Siewert 1998), and normativity (Brandom 1994). Again, the predominant view even among advocates of such explanatory gaps in psychology was that the psychological gaps stand in marked contrast to the natural sciences, which were generally viewed as a grand hierarchy connected by intertheoretic reductions. Chalmers (1996, 93), for example, writes, "Almost everything in the world can be explained in physical terms; it is natural to hope that consciousness can be explained in this way too. . . . However, I argue that consciousness escapes the net of reductive explanation."

The status of the explanatory gap in psychology is still controversial. Some view it merely as a symptom of the current immature state of the relevant sciences. Even among advocates of the gap, there is no consensus on whether even a principled and abiding *explanatory* gap has any metaphysical implications in the form of a failure of supervenience. On the one hand, Chalmers (1996) and Chalmers and Jackson (2001)[4] have followed Descartes in holding that metaphysical necessities must be, at rock bottom, conceptual necessities, and hence that a principled explanatory gap implies a failure of supervenience as well. They have thus opted to revive forms of dualism. Others, such as McGinn (1983) and Thomas Nagel (1986), have suggested the "Mysterian" view that there might be principled reasons, of a purely epistemological nature, why the mind might have special problems in completely understanding how it is necessitated by its own supervenience base, thus giving rise to unique explanatory gaps in psychology. Some Mysterians remain agnostic on the further question of whether the mind supervenes upon physical phenomena, while others are nonreductive materialists, holding that mind does supervene upon matter, albeit in a fashion unfathomable to the kinds of minds in question. Davidson (1970), at least, asserts token identity between mental and physical particulars, but due to his interpretivist view of mental-state ascriptions, denies that physical descriptions pick out unique mental descriptions. Searle (1992) seems to play both sides, denying the reducibility of mental states while asserting that mental states are causally necessitated by the biological nature of the brain.

2.5. Nonreductive Physicalism and Supervenience

Davidson's views also spawned a resurgence of interest in nonreductive forms of physicalism. Nonreductive physicalists claim that mental states in some sense *are* physical states, even though they cannot be explained by being reduced to physical states. While Mysterianism (the view that mental states cannot be explained in the same fashion as the phenomena of other special sciences) is one approach to nonreductive physicalism—one that addresses the explanatory gaps at an epistemological level—other nonreductive physicalisms attempt to address mind-body relations at a metaphysical level, treating them as contingent identities or even as necessary but epistemically opaque determination relations, or necessary identities underwritten by the New Semantic analyses of Kripke and Putnam.

Most nonreductive physicalists hold that mental phenomena are metaphysically supervenient upon physical phenomena, though some (like Davidson) deny that there are even psycho-physical laws. (Davidson claims that anomalism is consistent with supervenience of the mental upon the physical, but it is a peculiar form of supervenience, in that the physical base stands in the same relation to a number of mentalistic interpretations, with none holding pride of place.) There are many philosophical variations on the notion of supervenience. The basic idea behind supervenience is that, once one has fixed in place certain features of the world (what we might call "*basal* states"), certain other features of the world ("supervening states") are thereby fixed as well. A state S supervenes upon a state B just in case there cannot be a change in S-properties without a change in B-properties. Equivalently, if S supervenes upon B, then B→S is necessarily true. For physicalists, the B-features par excellence are phenomena of basic physics, and the most important type of S-features are mental phenomena (though some would hold that mental states supervene more directly upon neural states).

There are several variations on the notion of supervenience. These vary along two dimensions. One dimension consists in the *modal strength* of the dependency. A strong form of supervenience is *metaphysical* supervenience. S is metaphysically supervenient upon B just in case there are no two possible worlds that are exactly alike in their distribution of B properties yet are divergent in their S properties. A weaker notion of supervenience is called "*nomic* supervenience" or "*natural* supervenience." S is nomically supervenient on B just in case there are no two worlds *sharing the same natural laws* that are alike in their B properties but are divergent with respect to their S properties. A dualist can embrace nomic supervenience by holding that, in addition to the *physical* laws, our world also has psycho-physical laws relating physical states and mental states in a lawlike way (cf. Chalmers 1996). It is thus metaphysical supervenience that is needed to distinguish physicalism from dualism.

A second dimension defining different varieties of supervenience is based on how broadly the supervenience base is interpreted as being. A mental state is *locally* supervenient upon a physical state or brain state just in case some relatively local physical or neural facts—for example, facts about an individual's brain—fix the mental states as well. *Global* supervenience, by contrast, is the notion that things like mental states are fixed once one has fixed *all* of the physical facts for the entire world. Global supervenience is attractive in part because of the popularity of externalist views of intentional content—for example, that the fact that my concept GOLD refers to the element Au is not fully determined by what is in my head, but also depends upon my causal and ostensive relations to the environment. However, what externalism really demands is not so much a dependency upon *everything* about the world, but relatively local facts that are nonetheless not so local as to be confined to the being possessing mental properties. In my view, such considerations are better handled by way of a somewhat loosened notion of what counts as "local" than by recourse to all of the physical facts about the world. Retreating to global supervenience fails to demarcate the properties that really matter to the (supposed) supervenience relations, as the *same* supervenience base must be used to underwrite *all* supervening states.

2.6. Summary of Mainline Views

We may summarize these views by way of the following, locating each of these familiar positions in philosophy of mind with respect to their answers to the following questions:

1. *Reduction in the natural sciences*: Is intertheoretic reducibility the rule among the natural sciences?
2. *Supervenience in the natural sciences*: Do the phenomena of special sciences like chemistry and biology supervene upon physical facts?
3. *Psychological reduction:* Can mental phenomena like consciousness and intentionality be reduced to facts in the natural sciences?
4. *Psychological supervenience*: Do mental phenomena supervene upon the facts of the natural sciences?
5. *Positive Epistemology-to-Metaphysics Connection (Positive EMC)*: Does a reduction of A to B entail that A supervenes upon B?
6. *Negative Epistemology-to-Metaphysics Connection (Negative EMC)*: Does the irreducibility of A to B entail that A does not supervene upon B?
7. *Normative Reductionism*: Does the irreducibility of a phenomenon A to facts statable in terms of the natural sciences imperil the scientific and ontological legitimacy of A?

This summary of contemporary views in philosophy of mind highlights two sets of issues (see Table 2.1). On the one hand, there are the issues that are broadly contested: whether mind is reducible to material processes; if not, whether this implies a failure of supervenience as well; and whether reducibility acts as a kind of litmus for scientific and ontological legitimacy. On the other hand, there are also issues on which there is a broad consensus: whether a successful reduction

TABLE 2.1. Table of Mainline Positions in Philosophy of Mind

	Reduction in natural science	Supervenience in natural science	Psychological reduction	Psychological supervenience	Positive EMC	Negative EMC	Normative reduction
Reductionists	Yes	Yes	Yes	Yes	Yes	Yes	Yes/No[a]
Eliminativists	Yes	Yes	No	No[b]	Yes	Yes	Yes
Dualists	Yes	Yes	No	No	Yes	Yes	No
Mysterians/ Nonreductive Materialists	Yes	Yes	No	Yes (mostly[c])	Yes	No	No

[a]Some reductionists, like the Positivists, took reducibility to be a normative condition, while others, like Oppenheim and Putnam, took it only as a hypothesis.

[b]For eliminativists, the failure of supervenience is a trivial consequence of the claim that there are no mental states to thus supervene.

[c]Davidson and other interpretivists either reject supervenience on the grounds that there are always multiple equally good intentional characterizations of a person's behavior, or hold to an odd version of it in which a physical description does not imply a unique mental description.

assures supervenience, and whether reductions and supervenience are broadly to be found within the natural sciences. Advocates of the views canvassed have been fighting it out in the philosophical trenches over the former set of issues, and one way of proceeding is to enter into the fray on one side or another.

My intention, by contrast, is more subversive: to argue that almost everyone concerned has been *laboring under a mistaken assumption in thinking that reductive explanations are widespread within the natural sciences,* and that those who additionally believe that reducibility serves as a metatheoretical norm are mistaken on that account as well. If I am right in this, then almost everything is up for grabs, and philosophy of mind needs to go back to the drawing board and rethink some of its most familiar problems. This being the agenda for the book, I shall resist the temptation to enter into too many existing skirmishes and only say a bit more about the disputed issues insofar as it serves to illustrate the underlying importance of more general assumptions about reduction that I shall attack in later chapters.

2.7. Broad Reduction and Conceptual Adequacy

We have thus far characterized "broad reductionism" only very informally, saying that it is broader and weaker than the type-identity thesis, and that its core commitment is to bottom-up explanations. A bit more formally, the characteristic features of broadly reductive explanations are that they are:

1. *Part-whole explanations* (i.e., explanations of features of an entire system in terms of the properties and relations of its proper parts, or of elements lying at an ontological level no more complex than that of its proper parts), and
2. *Explanations without remainder,* or alternatively, conceptually adequate explanations.

The first part of this definition—that reductions are part-whole explanations—is for the most part self-explanatory. Chemistry deals with atoms and molecules, while particle physics deals with subatomic parts. A physical reduction of chemistry would require an explanation of chemical kinds, properties, and laws in terms of the properties and laws of subatomic particles. A reductive explanation of the mind would first have to take the mind to be composed of parts (say, networks of nerve cells), and then explain the properties of the mind in terms of the behavior of these parts.

Some might object to this definition on the grounds that one can be an *externalist* about things like mental content (i.e., believe that they are partially determined by things outside the thinker) and still be a reductionist. This is at least in part a dispute over terminology. If the "parts" invoked in a reductive explanation must be the parts of the very system that is being explained—for example, if they are confined to the parts of a single thinking organism—then externalist explanations are nonreductive, and indeed the need for externalist

explanations (not only in psychology, but also in biology and arguably in physics itself), might itself be invoked as an objection to reductionism. However, some reductionist writers, like Oppenheim and Putnam, have located the issue of part-whole relation not in the meriological relation between the reduced and reducing entities, but in the *level* of description. That is, the issue is not that a reductive explanation of A must be *cast in terms of relations of things that are parts of A itself*, but that it must explain A in terms of things *that lie at a simpler level of composition than A itself*. Thus a reduction of chemistry cannot be cast in terms of chemical or biological properties, a reduction of psychology cannot be cast in psychological or sociological terms, and so on. And so an externalist account to the effect that, say, the concept WATER refers to H_2O and not XYZ might still appeal only to the levels of organization that are simpler than those of cognizers, that is, whatever systems are needed to explain narrow content plus the causal relation to the molecular kind in question. I think that there are substantive issues at both levels. (It is harder, for example, to address social [Burge 1979] as opposed to causal [Putnam 1975] externalism in this manner.) For now, I leave open the possibility that a broadly reductive explanation of the mind may appeal to factors outside of the organism, so long as these are couched at a level of organization that is itself subpsychological.

I shall point out, however, that accounts that appeal to the state of the *whole* physical world to explain mental properties (or those of the special natural sciences) would save only global *supervenience* and not reductive *explanation* of mental types. Chalmers (1996), for example, suggests that all nonmental phenomena are metaphysically supervenient upon a perfectly global physical description of the world. But even if this is so, it does not entail that they are reductively *explainable* in physical terms. Such an explanation would require us to specify types of physical description (albeit relational ones) that would underwrite the applicability of descriptions in the special sciences. This requires finding a level of explanation that is neither local to the phenomena being explained nor completely global, in the form of a list of all of the physical facts. Externalist theories of mind generally *do* at least implicitly specify such a level; for example, the organism in its causal and normative relations to physical objects and social/linguistic practice. However, for such a level of explanation to be able to provide the basis for a *reductive* explanation of the mental, the things it appeals to must be specifiable without recourse to things like consciousness, intentionality, and normativity. It strikes me as dubious that this can be done for *social* externalism, which generally appeals to intrinsically intentional and normative phenomena such as language. Whether it can be done for *causal* externalism depends on (a) whether causal externalism is sufficient *without* social externalism, and (b) whether objects, kinds, and properties can be carved into kinds in a canonical way that is mind-independent. This latter issue I shall put on hold for the moment. But the Cognitive Pluralist account offered in part III implies a rejection of this degree of realism about objects, kinds, and properties.

Another potential ambiguity must also be avoided. Scientists are often wont to use the word 'reduction' quite generically for *any* part-whole explanation, however incomplete. Thus for example, we hear of the discovery of the

efficacy of Prozac and other selective serotonin reuptake inhibitors in providing a "reduction" of depression by means of properties of the serotonin reuptake mechanisms in the nervous system. Such "reductions" often fall short, not only of explaining the subjective properties that are the locus of the explanatory gaps, but even of providing full mechanisms to explain the nonphenomenological data. A slightly stronger usage is suggested by Bickle (1998) in his notion of "new wave" reductionism, which treats the identification of elements from different domains (e.g., pains and C-fiber firings) as *contingent* rather than a necessary consequence of part-whole relations. While I applaud attempts to explicate actual scientific usage and hope that this sort of hands-on philosophy of science will flourish and grow, it is important to recognize that so weak a notion of "reduction" would deprive us of any connection between reducibility and any metaphysical result (a result actually embraced by Bickle 2003). To explain facts about atoms by appeal to strong force, for example, does not entail that *all* atomic-level properties supervene upon strong force interactions between their constitutive particles. (Forces like weak force, gravitation, and electromagnetism are also at work.) Philosophers have traditionally been interested in a notion of reduction that is closely connected with metaphysical necessity, and with good reason: if you have a part-whole explanation without remainder of A in terms of B, you have thereby guaranteed that B→A is metaphysically necessary as well. Weaker notions of "reduction" do not underwrite metaphysical supervenience claims. A weaker notion of "reduction" would also render it incapable of underwriting the unity of science, at least in Carnap's sense of uniting the sciences in the form of a single grand axiomatic reconstruction.[5] Since connections with metaphysics and unification have historically been important parts of the reductionist project in philosophy of science and philosophy of mind, I shall here reserve the word 'reduction' for a usage that implies not only part-whole explanation, but part-whole explanations that are also comprehensive, in the sense of explaining *everything* about the reduced system.

I have elsewhere referred to such explanations as "conceptually adequate explanations" (CAEs): "An explanation of *A* in terms of *B* is conceptually adequate just in case the conceptual content of *B* is sufficiently rich to generate that of *A* without the addition of anything fundamentally new" (Horst 1996, 267). CAEs, moreover, come in a variety of strengths. The strongest form, which we may call a *Pure CAE*, is found in syllogisms and mathematical constructions, in which all of the conceptual content of what is explained is either already present in, or else constructed from, the conceptual content of the explaining system. In a logical syllogism, all of the predicates in the conclusion must be present in at least one of the premises. In a locus-theoretic definition of a geometric figure, the figure's definition is constructed out of more primitive definitions (e.g., a circle is defined as the locus of points equidistant from a single point on a plane).

A slightly weaker form of explanation is obtained when we augment the vocabulary of the explaining system with additional, but purely formal tools, such as Descartes's application of algebraic tools to Euclidean geometry to create analytic geometry. Such additions are metaphysically innocuous, as they

confer additional inferential power without the risk of sneaking in additional ontology or substantive properties in the process. Let us call this form of explanation *Formally Augmented CAE*.

There is at least one additional, and still weaker, form of CAE. Consider the explanation of how a given token system counts as a member of a functional kind: how a bit of tissue counts as a heart, or how an arrangement of circuits counts as an AND-gate. In this case, the nature of the *kind* ("heart," "AND-gate") is not explained in the process. The anatomy of the heart does not explain *what it is to be a heart*. However, if the functional kind is suitably well-understood, it becomes epistemically transparent how a system having *this* set of *physical* properties would thereby have *that* set of *functional* properties as well. The nature of the kind lays down, as it were, criteria for what could count as a member of that kind, and so this form of explanation might aptly be designated *Criterion-Filling*. Unlike constructions, Criterion-Fillings do not supply reductions of the kinds themselves, but may (when they are part-whole explanations) supply reductive explanations of a particular object's being a member of such a kind. (It is possible that there are kinds other than functional kinds that can be explained in this way as well.)

The Rationalist and Positivist traditions sought primarily to model their explanations on things like syllogisms and constructions, which is to say on Pure or Formally Augmented CAEs (though their success in this was arguably compromised by the need for metaphysically contingent bridge laws). As Chalmers (1996) points out in different language, current optimism about reduction in philosophy of mind tends to trade on the assumption that psychological kinds are functional kinds, suggesting that these are Criterion-Fillings.

2.8. Broadly Reductive Naturalism: Some Theses

We have noted that broadly reductive views of the mind embrace several different types of claims. On the one hand, some are about reduction as a form of *explanation*; others are really claims about *metaphysics*. On the other hand, claims that the mind can be naturalized are sometimes posed as positive, or second-order, empirical claims, but are sometimes posed as normative claims as well. It is useful at this point to distinguish a number of claims that are variously made by reductive naturalists, organized along these two axes.

2.8.1. Claims about Explanation

2.8.1.1. Positive Claims about Explanation. These range from the very mild EP 1 to the extreme seventeenth-century rationalist views in EP 8 and EP 9, which, to the best of my knowledge, are universally rejected today.

EP 1: Broadly reductive explanations have played an important role in the natural sciences.

EP 2: There are successful broadly reductive explanations uniting scientific domains.

EP 3: Broadly reductive explanation is (in fact) the only form of intertheoretic explanation employed in the natural sciences.
EP 4: There is a "basic" science to which all other "special" sciences can be reduced by broadly reductive explanation, that is, such that all more complex phenomena can be constructed from it (at least in principle) through the kind of axiomatic reconstruction undertaken by E. Nagel (1961) and his followers.
EP 5: This basic science is basic physics.
EP 6: Given a complete physical description of the world at time T, one could (in principle) derive *all* facts about the world at T ("vertical" prediction).
EP 7: Given a complete description of the world at T, one could (in principle) predict all subsequent events at T + n (Laplacean or "horizontal" prediction).
(EP 8): For any complex phenomenon P, there is a *unique* (correct) resolution of P to the basic science (typical only of some versions of the view, and generally rejected since the functionalist movement).
(EP 9): The basic science is knowable indubitably (typical only of seventeenth-century versions of the view).

2.8.1.2. Epistemic/Normative Versions

EN 1: Broadly reductive explanation is the only *legitimate* form of intertheoretic explanation.
EN 2: For any special science S, either S can be reduced to basic physics through broadly reductive explanation, or else the legitimacy of S is undercut (and with it our warrant for believing in the objects postulated in S).

2.8.2. Claims about Metaphysics

2.8.2.1. Positive Claims

MP 1:There are simple and composite facts. (Substitute 'objects' or 'properties' for 'facts' to get different variations.)
MP 2: All complex facts are determined by simple facts.
MP 3: The only simple facts are those of basic physics.
MP 4: Mental facts are not facts of basic physics.
MP 5: Mental facts are determined by physical facts, and ultimately by basic physics; that is, given a suitably complete set of basic physical facts at time T, these determine a unique possible mental state at T as well (vertical determination; implied by MP 2, MP 3, MP 4).
MP 6: Mental facts are just complex physical facts.
MP 7: Given a suitably complete set of physical facts at T, these determine what mental states will take place at all subsequent T + n (horizontal determination).

2.8.2.2. Metaphysical/Normative Claim

MN 1: If mental states (events, properties facts) are not determined by physical facts, the ontological legitimacy of the former are imperiled.

2.9. Reduction and Metaphysics: Some Historical Notes on "Math Envy"

It is worthy of note that reductionism enjoyed its greatest popularity in two periods: the seventeenth and the twentieth centuries. In both cases, it was developed explicitly on the model of mathematical reasoning. For early moderns advocates like Galileo, Hobbes, and Descartes, the paradigm to be emulated was geometric construction and proof. For the Positivists and Logical Empiricists, it was the logical syllogism and axiomatic systems. In mathematics, a rich set of theorems and constructions can be developed out of a much smaller set of primitive definitions and axioms. Reductionists from Hobbes and Descartes to Carnap and Ernest Nagel supposed that complicated physical systems could likewise be understood as derivations or constructions out of characterizations (definitions) and laws (axioms) pertaining to the simplest material bodies. (Descartes, of course, drew the line at language and reason, but the physics of both *Le Monde* and the *Principles* was distinctly reductionistic.) The early moderns saw this form of reasoning as constitutive of *scientia*. In the philosophical usage of the day, rigorous knowledge (*scientia*) by definition had to take the form of things that were known either directly and self-evidently, or else by way of transparently valid deductions from things known in the first way. So, if there was to be a *scientia* of nature, it would have to be like mathematics in its form. Platonists and Aristotelians had been of the opinion that such exact knowledge of the material world was not possible.[6] Much of the allure of the Cartesian Method and Galileo's Method of Resolution and Composition lay in the hope each extended for such a natural *scientia*.

A second important appeal of the reductive approach has always been that it forges an immediate connection between explanation and metaphysics. If you can reduce A to B, you get metaphysical supervenience for free. If A can be derived from B, then B→A is true in every possible world. As discussed in chapter 1, this virtue stands in stark contrast to what one obtains from other important forms of explanation, even ones that have enjoyed prominent roles in the natural sciences, and it is arguably for this reason that reductive forms of naturalism have historically enjoyed more fame than other forms among philosophers. Newtonian or nomic explanation does not underwrite metaphysical supervenience, and evolutionary explanation seems to stand or fall with phenotypic features being a necessary consequence of facts about genes plus development plus environment.

It is also important to note that, historically, it was the plausibility of *reductionism* that drove the plausibility of *physicalism*, and not the other way around. At the time of Galileo and Descartes, the prevailing view of the world

was Aristotelian and had a wildly profligate ontology, in which there were separate types of substance for each animal and vegetable species. The predominant form of Aristotelian scientific explanation ("natural motion") appealed to the "specific nature" of each thing that was to be explained (e.g., to explain the spider's web, you appeal to the role webs play in the form of life characteristic of that species of animal—its "nature"). Part of the power of the mechanist alternative was that it could provide better explanations of a great many phenomena by postulating just one type of substance ("matter" or "body") and one type of change ("[local] motion," in the sense of change of position). It was the apparent availability of this type of explanation that rendered plausible Hobbes's materialism and allowed even Descartes to postulate that everything except God and certain properties of souls (rationality, language, voluntary action, first-person experience) could be explained by, and consist in nothing over and above, the motions of tiny material bodies.

In short, reduction is not only a knock-down argument for metaphysical supervenience, at least when you can get it, but it also seems to be the *only* way to provide such a knock-down argument: it is hard to see just how one could motivate supervenience or physicalism nearly so compellingly in the absence of a conviction of reducibility. Thus, to the extent that one were to reject reductionism, one would need to reconsider one's grounding for assuming supervenience and physicalism as well.

2.10. The Empirical Status of the Psychological Gaps

While the explanatory gaps in psychology enjoy wider credence today than they did a few decades ago, their status is by no means uncontroversial. In particular, arguments for *principled* gaps are generally *philosophical* arguments, and philosophical arguments have a nasty habit of being trumped at a later date by empirical discoveries. (One thinks, for example, of some of the Scholastic objections that Copernicanism was literally nonsensical; that is, it was nonsensical *by Aristotelian/Ptolemaic lights*. Or of objections from Leibniz and other mechanists that Newton's gravitational theory was nonsensical because it required "occult forces" of action at a distance.) I do thus take arguments for the psychological gap to be *presumptive* or *prima facie* arguments that are vulnerable to criticism on empirical grounds. However, as someone who has studied several notable forms of explanation in the sciences of cognition, I would tend to say that thus far, at least, they tend to confirm that there are indeed explanatory gaps exactly where the philosophers predict them to be (Horst 2005).

This claim is *not* intended to be antiscientific. Rather, it is based on careful consideration on just what is and is not explained in various types of explanation in the cognitive sciences. It would be distracting from the main task of this chapter to explore examples in detail, but it might be useful at least to mention some that are representative. Localization studies, marshaling evidence from trauma cases, brain imaging, single-neuron sampling, and animal experiments,

do lead to conclusions about what parts of the brain are implicated in some special way in the exercise of various psychological capabilities. But this does nothing to explain the fact that their related psychological states have a phenomenology at all; nor does it decide the question of the metaphysical status of these special connections for example, whether they are instances of identity, supervenience, or mind-brain causation. Psychophysical laws like the Weber laws express regularities in the relations between stimuli and percepts, but do not explain why the percept has the precise qualitative character it has, nor indeed why it has any qualitative character at all. Models of early color vision explain things like metamers, the three-color process postulated by Helmholtz, color opponency, and the shape of the human color solid. But again, they do not explain why any of this involves a phenomenology, or the specific phenomenology it has (Horst 2005). Viewing the mind as a syntactically driven symbol-processor at most explains mechanisms underlying reasoning and the generativity of thought, but it does not explain meaning (Horst 1996). Causal theories of meaning at most explain the meaning-*assignments* of mental states—why *this* type of mental state refers to *that* thing—but they do not explain how meaningfulness gets into the picture in the first place, as causation is too bland and ecumenical a relation for that (Horst 1996).

The list could be expanded almost indefinitely, but the reader will see the general point. I shall thus assume, for purposes of this book, that the thesis that there are psychological gaps has not been refuted by the sciences of cognition, even though it is in principle subject to such refutation.

2.11. Negative EMC

The status of the Negative Epistemology-to-Metaphysics Connection principle (Negative EMC) is likewise controversial. I see the current state of play as follows. There is an intuitively compelling argument in favor of the principle, which presents a *prima facie* or presumptive case for it. This, however, has had to endure several important challenges in the form of potential defeaters for the argument. The basic argument is quite simple and elegant. It has been made in different terms by a number of authors, often appealing to a notion called "conceptual necessity." I prefer, however, to phrase it in terms of conceptually adequate explanation (CAE).[7]

A1. There are mental properties M for which there is no set of nonmental properties N such that it is possible to give a CAE of M in terms of N (*the Explanatory Gap.*).

A2. If it is impossible to give a conceptually adequate explanation of A in terms of B, then it is not metaphysically necessary that B implies A (*Negative EMC*).

A3. There are mental properties M for which there is no set of nonmental properties N such that it is metaphysically necessary that N implies M.

I shall treat A1 as granted for purposes of argument. The second premise is where the main burden of the argument rests. I hope that the reader will agree with me that even the A version of this argument holds a great deal of intuitive force.

However, the key premise here, A2, is by no means uncontroversial, and in fact touches on issues that have excited major interest in metaphysics and philosophy of language over the past quarter century. The first objection is that the Main Argument assumes a sort of descriptivist semantics that has been largely overthrown by Kripke (1972) and Putnam (1975). While neither Kripke nor Putnam is a proponent of reductionism, many have viewed their "New Semantics" as supplying a refutation of the Negative EMC. I shall argue that this view is mistaken. The second objection stems from the availability of nonreductive versions of physicalism that treat mental properties as "emergent"—that is, as supervenient upon but not reducible to the physical.

2.11.1 The Objection from New Semantics

The Negative EMC asserts a strong connection between the *epistemological* relationship of *conceptual entailment* involved in CAEs and the metaphysical relation of *metaphysical necessity*. However, this connection between rational implication and metaphysical determination is precisely one of the things that the New Semantics calls into question. On the one hand, there are necessity claims that are true, but not derivable on the basis of the sense of the terms. If B is "x is water" and A is "x contains hydrogen," then B→A is necessarily true, even though one could be a competent speaker of English and not be able to tell, on the basis of the sense of 'water', that 'water' refers to H_2O, and hence necessarily contains hydrogen. On the other hand, there are necessity claims that are underwritten by the sense of the terms and yet turn out to be false. For example, before Aristotle, it may well have been one of the constitutive implications of (the Greek equivalent of) the word 'whale' that "x is a whale" implied "x is a fish." One might imagine Greek philosophers using (the Greek equivalent of) the sentence "All whales are fish" as an example of an analytic truth, much as we might use "All tuna are fish" as such an example today. They would, however, have been mistaken.

If words like 'pain' and 'thought' work like the paradigm instances of "natural kind terms" like 'whale' and 'water', then it is not safe to draw metaphysical conclusions on the basis of the senses of those terms either. On the one hand, sentences like "All pains are C-fiber firings" could be necessarily true, even though underivable from the sense of the constituent expressions. On the other hand, purported necessary truths like "Mental states do not have extension" could turn out to be empirically false if the referents of the expression 'mental state' turn out to be, say, the members of a class of brain states. If 'pain' means something like "the kind of event, whatever it is, that is the natural kind underlying this nasty sensation," and the kind of event in question turns out to be C-fiber firings, then what 'pain' picks out are C-fiber firings. Indeed, "Pains are C-fiber firings" would then prove to be necessarily true, and *to have been necessarily true all along*, albeit only in the rather uninteresting sense that *it turns out that* the word 'pains' in fact *refers* to 'C-fiber firings',

and so the truth conditions of 'pains are C-fiber firings' (read *de re*) are the same as those of 'C-fiber firings are C-fiber firings'. The truth is *necessary* because it states an identity, and identities are always necessary.

Of course the *de dicto* reading of "Pains are C-fiber firings" is not necessarily true unless it is true, in every possible world, that what causes the relevant sort of nasty sensation is a C-fiber firing. Hence C-fiber firings need not be the only things that could produce nasty sensations, any more than H_2O is the only kind of stuff that can, in sufficient quantities, be a clear, nonviscous fluid found in lakes and streams. The identity theorist does not need to claim anything so bold. Indeed, she may well say that nasty sensations are *likely* to be realized differently even in other actual species, since they have different nervous systems. What we are to take from the analogy, rather, is a model for seeing how the identification of pains with C-fiber firings could be *true*, even though it is not a truth of analysis. One condition for its being true in this way is that the reference of 'pain' is fixed in the same way that the reference of 'water' is allegedly fixed: that is, by a causal or ostensive relation formed by saying, as it were, " 'Water' shall refer to *that* stuff, whatever it turns out to be, which is a clear fluid found in lakes," or " 'Pain' shall refer to that kind of event, whatever it turns out to be, that produces *that* kind of nasty sensation." Here one is using the sense of the term to help pick out a kind of thing, but the term is then supposed to track the kind of thing picked out, and not the criteria employed to pick it out: H_2O rather than clear liquids, C-fiber firings rather than nasty sensations.

Of course, even at its best, this account may seem a frail reed for the physicalist to hang his or her hat on. It does nothing to show that the mentalistic vocabulary *does* pick out physical properties. But we are not examining it here in the role of an argument for physicalism, but as an objection to A2. And if mentalistic terms *do* work in this way, then A2 may be in trouble.

The remainder of this section addresses two questions:

1. For those terms to which New Semantics (NS) is applicable, does it really drive a wedge between metaphysical necessity and conceptually adequate explanation?
2. Are mentalistic terms like 'pain' and 'belief' ones to which NS is applicable?

Answering the first question requires a deeper examination of what is really going on in the NS analysis of the 'water' case. One of the puzzling facts about discussions of the implications of NS for philosophy of mind is that different people have drawn such different conclusions about what those implications are. (Kripke, after all, uses it in an argument *against* physicalism.) I shall suggest that there are in fact alternative ways one might *interpret* NS, and that these lead to rather different conclusions about the mind.

2.11.2. New Semantics and Conceptually Adequate Explanation

There is a crucial unclarity in the story that New Semantics is supposed to tell about the semantics of natural kind terms. On one interpretation of the story,

the inability to draw conclusions like "Necessarily, water contains hydrogen" or "Necessarily, whales are mammals" is a consequence merely of the sense of the word 'water' in current usage, whether in ordinary language or in scientific theory. The inability of prebiological peoples to see that whales are necessarily mammals is merely a symptom of their ignorance about whales. This ignorance vanishes when a better biology comes along, and with it the opacity of the inference. Likewise, a modern chemical understanding of water would allow one to *see* that "Water contains hydrogen" is a necessary truth.

On the other interpretation of the story, there is a deep and abiding epistemic opacity to the inference in question. On this interpretation, *not even an ideally complete* scientific understanding of water or whales would allow one to see that the truths in question are necessary truths. I shall refer to these interpretations as *canonical essentialism* and *opaque essentialism*, respectively.

> *Canonical Essentialism*: Natural kind terms pick out essential properties of mind-independent natural kinds, and one could in principle arrive at a "canonical (re)formulation" of the sense of any concept whose constitutive inferences would parallel all of the metaphysically necessary determination relations of the property.
>
> *Opaque Essentialism*: Natural kind terms pick out essential properties of mind-independent natural kinds, but (at least in some cases) their nature is to some extent *epistemically opaque*, so that there is not a canonical (re)formulation available wherein the constitutive inferences would parallel all of the metaphysically necessary determination relations of the property.[8]

The canonical essentialist interpretation cannot be used as a defeater for Negative EMC. Indeed, the canonical essentialist *embraces* Negative EMC. She simply holds the entirely reasonable view that the entailment of such a necessity claim need not be evident on the basis of ordinary-language semantics or that of an inadequate scientific theory. But this is not inconsistent with Negative EMC. Negative EMC claims only that the inavailability-*in-principle* of a reduction implies a failure of supervenience. It does not claim that supervenience is imperiled by the mere lack of a conceptual entailment in ordinary language or on the basis of an inadequate theory. The question is whether the present lack of a reductive explanation of pain and other mental phenomena is a result of lacking the sort of more adequate understanding that we now have in the case of water or whales. And this question is left untouched by NS. Kripke, Jackson, and Chalmers argue that the inavailability of a reductive explanation of *mental* properties is *not* simply a matter of ignorance, but rather that the physical sciences do not (and cannot) supply even any *candidate reducers* for things like consciousness or intentionality. The burden of proof here would seem to be on the would-be reducer to produce the kind of more adequate canonical descriptions that would make such a reductive explanation possible. Until then, this position is merely an expression of a faith or ideology.

The opaque essentialist position is internally consistent and is incompatible with Negative EMC; but it is problematic as well. For it would plainly *not* seem to be what we find in the standard examples of supposed *a posteriori* necessity. We *do* have understandings of water and of whales that permit us to derive claims of metaphysical necessity, given the assumption that these terms function as NS claims natural kind terms to function. Adopting a nonreductive physicalist position would require us to postulate a crucial asymmetry between kind-terms like 'water' and 'whale' on the one hand and 'pain' and 'judgment' on the other: a proper understanding of water or whales *does* provide the basis for a CAE of water's containing hydrogen or whales' being mammals, while there seems to be no understanding of the brain that licenses similar inferences about qualia or intentionality (e.g., understanding water as the chemical compound H_2O analytically entails that it contains hydrogen). It is possible to take the view, with McGinn and Nagel, that there is a *special* opacity encountered when the mind attempts to understand itself. However, this view owes nothing to NS, as it is not a general consequence of the NS analysis of kind-terms, and it requires a separate motivation.

2.11.3. The Objection from Mysterianism and Nonreductive Physicalism

To reject Negative EMC on Mysterian or nonreductive physicalist grounds requires one to hold two separate views. The first is that the mental is metaphysically supervenient upon the physical. The second is that one cannot reductively explain the mental in terms of the physical. The first is a claim about metaphysics, the second a claim about explanation. I shall stipulatively use the term 'emergentism' to refer to this combination of views. 'Emergence' in this sense is defined as (abidingly) epistemically opaque metaphysical supervenience. It is defined in contrast with 'resultant' properties, which are defined as properties that are metaphysically supervenient in a fashion that can be reductively explained, and with nonsupervenience. (Nonreductive physicalism is not limited to emergentist views. Davidson's anomalous monism, for example, involves the claim of token identity without the further view that there are "psychophysical laws" associating a physical state with a unique mental state, on the grounds that Davidson believes that there are multiple equally good interpretive assignments of mental states in any given instance.)

Emergentism strikes me as a *consistent* position so long as one does not also embrace a fairly strong version of rationalism that holds that everything about the universe is intelligible to minds like ours. I tend to think that the Mysterian side of emergentism is wholly reasonable. I see no reason to think that either God or evolution would endow human minds with the capacity to understand the real essences of everything, even if one believes there are such essences. The problem with emergentism lies in the fact that it embraces physicalism while rejecting reduction, which is plausibly the only basis on which physicalism could be shown to be true. The problem is thus one of the evidential and rhetorical status of the physicalist component of emergentism.

The appeal of emergentism would seem to depend crucially on how one arrives at it. It is my impression that most emergentists are disenchanted

reductionists. They aspired to a reductive physicalism, but then became convinced that there are in fact principled and abiding explanatory gaps. From this standpoint, emergentism seems like an appealing and minimal tactical retreat. But consider how emergentism looks if one approaches it from the opposite direction. Suppose one starts out as a Mysterian and then poses the question, "What is the best metaphysical interpretation of the explanatory gaps we find in the case of the mind?" Indeed, suppose that we concede to such a person the claim that there are robust empirical generalizations relating mental states with brain states. (Such a concession may be problematic in some cases, such as free will, but it may be comparatively innocent in cases like sensory psychophysics.) We then ask whether such generalizations are best seen as the products of epistemically opaque metaphysical necessities, or of something weaker, such as causal laws. What is to decide between such interpretations?

Note that here we have moved beyond the boundaries of empirical science and into metaphysical speculation. Given that there are robust empirical generalizations about mind-body relation, there is no further empirical test that can decide between modal interpretations with stronger force than empirical generalization. Considerations of parsimony are irrelevant here. Occam's Razor can at most be invoked when we are dealing with two competing theoretical explanations. But in the case of the explanatory gaps, we have no such explanations to compare at the level of scientific theory, but merely competing modal interpretations. Occam's Razor is not intended to guide us in comparing two *non*explanations. Deciding on a necessitarian interpretation may be a *consistent* option, but it is ultimately a decision based on factors other than evidence. One may embrace an emergentist physicalism as a matter of taste, but one cannot claim the high ground of scientific progress.

2.11.4. Does the New Semantic Analysis Apply to Mental State Terms?

I am additionally skeptical about the assumption that mental states like pain and belief are among the things to which the NS analysis is properly applied. Kripke and Putnam's analyses are not intended to apply to *all* referring terms, but only to a certain sort. An examination of standard developments of such a semantic analysis reveals what kinds of terms we are dealing with. The sense of 'water' is supposed to be something on the order of "the natural kind, whatever it may be, of *that* potable, clear stuff found in lakes and rivers." Thus construed, 'water' is what we might call a *role-filler term*. The analysis picks out a *role*, constituted by the sense ("the stuff, whatever it is, that . . . "). The filler of that role, however, is underdetermined by the sense (the identifying criteria or constitutive implications), depending crucially on some further relation of causation or ostension (*that* stuff), and hence Earthlings and Twin-Earthers pick out different referents due to being in causal or ostensive relations to different types of stuff that fill the water-role.

For a similar analysis to apply to terms like 'pain' or 'belief', these terms would also have to be role-filler terms. But it is not at all clear that this is so. First, not all terms are role-filler terms. Consider Putnam's distinction between

disease terms and syndrome terms. Syndrome terms, such as 'headache', pick out a cluster of symptoms, while disease terms, such as 'polio', pick out a particular cause of those symptoms, though not necessarily the only thing that is capable of causing those symptoms. And there are familiar marks that differentiate how we proceed in problem cases with things like disease terms and how we proceed with things like syndrome terms. On the one hand, suppose doctors had assumed that a patient had polio, but then discovered that his symptoms were not caused by the polio bacillus. They would conclude that it was not a case of polio, and not that there was more than one cause for polio. 'Polio' is a disease term (and, more broadly, it is a *filler* term). On the other hand, suppose doctors had assumed that there was a unique cause for headaches, only to discover that a patient reported headaches, but that they were caused by a second underlying condition. They would *not* conclude that the patient did not have a headache after all; rather, they would conclude that there was more than one way of getting a headache. 'Headache' tracks a symptom or syndrome, not an underlying cause. That is, whereas 'polio' tracks the filler of a role, 'headache' tracks a role. Likewise, even if we allow that 'water' tracks the filler of a role, there are other terms (perhaps 'liquid') that track roles rather than their fillers.[9]

The crucial question, then, is whether terms like 'pain' and 'belief' are role terms or filler terms, or perhaps neither. For NS to apply, it is necessary that they be filler terms. But it is not at all clear that this is the case. Consider what one would say in the following cases:

a) There are painful sensations without C-fiber firings.
b) There are C-fiber firings without painful sensations.

Faced with case (a), we would say that pain was taking place; in (b) we would say that it was not. 'Pain' seems either to track role rather than filler, or to not be a role-filler term at all. (If your usage of the words is different, that is no problem: we can use the word 'pain†' for your usage and 'pain*' for mine. I'm willing to grant that [thus stipulatively defined] 'pain†' is a filler term. But the point is that there is also a perfectly good referring term, 'pain*' as we have now defined it, that does not work that way. And one is enough to make the point.)

Now the physicalist advocate of NS might object that such conceptual analysis is just the sort of thing that NS is supposed to make us suspicious of. And it is true that this kind of conceptual analysis about ''what we would say'' is not evidence of anything deep about the world. But we are not asking deep questions about the world here—we are asking a question about how we use a particular word: Is 'pain' the type of word whose reference functions like that of 'polio' or that of 'headache'? Examining our intuitions about counterfactual situations *is* the technique we standardly use to explore such questions; indeed, it is exactly the technique that is always used in developing things like Kripke's and Putnam's story about 'water' and other putative natural-kind terms. For example, we might consider what we would say in the following cases:

c) There is a clear, potable liquid found in lakes that is not H_2O.
d) There is H_2O that is not clear, potable or found in a lake.

In case (c), we would say that the liquid in question was not water. In case (d) we would say that it was water, but perhaps contaminated in some way. In both cases, 'water' tracks the filler and not the role, confirming the Kripkean analysis of 'water', but in contrast to the way the examples fall with 'pain'. The clear implication is that 'pain' requires a different sort of analysis.

What is characteristic of the broader class of filler terms is the use of something like ostension or causation in addition to sense or descriptive content to determine reference, and the fact that the reference tracks the object of ostension rather than the sense. But this is precisely what is *not* the case with 'pain' and other qualitative terms. ''Water that is not potable'' contains no contradiction, because the referent of 'water' is the ostended stuff and not *whatever* meets the identifying conditions used to pick it out on the occasion of ostension (some part of the sense). But ''pain that does not hurt'' is self-contradictory, because the reference of 'pain' tracks things that feel a certain way; that is, it tracks the identifying conditions themselves. In this sense, it is more like a role term than a filler term.

Kripke seems to have had similar concerns in mind when he used his analysis as the basis for a refutation of mind-brain identity. In *Naming and Necessity* he writes:

> Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus pain, unlike heat, is not only rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent. Thus it is not possible to say that although pain is necessarily identical with a certain physical state, a certain phenomenon can be picked out in the same way we pick out pain without being correlated with that physical state. (1972, 152–53)

Here, Kripke interprets 'pain' as picking out a property—that of feeling a certain way—essentially rather than accidentally, and contrasting it with the paradigm examples that license *a posteriori* necessities. The plain implication would seem to be that Kripke did not regard 'pain' as a filler term, and hence not a natural-kind term like 'water'. The latter picks out its referents ''accidentally'' and hence leaves open the door to *a posteriori* necessity. But because 'pain' is not a filler term, it picks out its referents essentially, and hence the door is not left open for *a posteriori* necessity to sneak in.

2.12. The Current State of Play

I would thus describe the current state of play in the following way. The explanatory gap (or multiple explanatory gaps for different mental phenomena

like consciousness, intentionality, and normativity) has raised significant *prima facie* problems for broad reductionism as a claim that one can *explain* mental phenomena like consciousness and intentionality in a particularly strong fashion. The fact that there at least appear to be psychological gaps suggests that the mind is not reducible to physical phenomena, *even if* reductionism is the rule in the natural sciences. Reductionism is not entirely refuted, however, since reductionists might (a) hold that the explanatory gaps are merely a symptom of current ignorance, or (b) hold that reduction is a metascientific norm that trumps the appearance of irreducibility. Proponents of eliminativism, a view that is in many ways the flip side of reductionism, might take such a norm to imply that what the explanatory gaps show is that there is a problem with the commonsense ontology of mental states. The existence of explanatory gaps is thus compatible with both eliminativism and nonreductive materialism. However, it is not clear how much evidence is left for the materialism once the reductionism has been cast to the wind.

Additionally, the Negative EMC presents an intuitively plausible *prima facie* case that explanatory gaps imply failures of metaphysical supervenience as well. This seems to imperil physicalism and support dualism. However, Negative EMC is highly controversial and involves us in deep metaphysical waters. Philosophers on both sides of the issue have taken its resolution to require new work in the semantics of modal claims (Chalmers and Jackson 2001; Block and Stalnaker 1999).

2.13. The Hidden Assumption

The backdrop for all of these positions is the assumption that reduction and supervenience are the norm *outside* of psychology. It is this assumption that makes the appearance that there are explanatory gaps in the case of psychology seem unacceptable to reductionists and eliminativists, and seem like a unique and sexy problem to dualists, Mysterians, and nonreductive physicalists. It is thus the linchpin that holds together an entire problematic in philosophy of mind today. It is also, by and large, an unexamined assumption within philosophy of mind.

It behooves us not to leave this assumption unexamined. First, should it prove false, a great deal of what is routinely assumed in philosophy of mind will have to be rethought. Second, we have already encountered one reason to examine it more closely, in the Mysterian suggestion that there are things that the mind is ill-suited to understand. If the mind is not suited to understand *one* thing, there may be *other* areas where it falls short as well: explanatory gaps may be more widespread than we imagined. Third, reductionism has in fact endured withering critiques within philosophy of science itself. Indeed, it would be hard to find many philosophers of science today who hold out much hope for Carnap- or Nagel-style reductions in the natural sciences. This is the topic of the next chapter.

3

The Demise of Reductionism in Philosophy of Science

The previous chapter surveyed a number of familiar views in philosophy of mind. All of these were predicated upon the assumption that intertheoretic reductions are widespread at the junctions between the various natural sciences and are available in principle at all such junctions. Some of those surveyed, moreover, held the additional view that reducibility serves as a kind of litmus or normative constraint on the scientific and even ontological legitimacy of special sciences like psychology.

These, however, are substantive claims within the philosophy of science. They were mainstream claims through much of the twentieth century and enjoyed their heyday in the middle part of that century. Curiously, though, at the very time that reductionism was reaching its zenith in philosophy of mind, these assumptions were coming under fire within philosophy of science itself. Today, these assumptions are in fact generally rejected by philosophers of science. In philosophy of science, the aprioristic normative agenda of the Positivists has been abandoned in favor of approaches that study the various methods and models of individual sciences, and the prevailing view is that the special sciences are autonomous and not in need of vindication by proving their reducibility to physics. The earlier optimism that the special sciences would prove reducible to physics has turned out to be largely unfounded. Reductions, in the relevant sense of that word, have proven few and far between, not only in the human sciences but in the physical sciences as well. And yet philosophy of mind has continued to labor under the yoke of an outdated philosophy of science. Indeed, it might not be an overstatement to say that turn-of-the-millennium philosophy of mind is one of the last bastions of

1950s philosophy of science. (It is, alas, not the only, nor even the best-known such bastion. Reading well-compensated, mass-market books like E. O. Wilson's *Conciliance* [1998] or Francis Crick's *The Astonishing Hypothesis* [1993], it is hard to escape the impression that the authors have not read any philosophy of science written since 1960.)

The burden of this chapter is to outline important post-reductionist developments in the philosophy of science. The remaining chapters assess their significance for problems in philosophy of mind.

3.1. "Naturalistic" Philosophy of Science and the Rejection of A Priori Influences

We have already seen that the Positivists and Logical Empiricists were fond of the idea of intertheoretic reductions in the middle part of the twentieth century. Their own reasons for favoring this metatheoretical view were rooted in philosophical concerns about the "logical form" of scientific explanations. The most important sea change in philosophy of science since the 1950s has been a rejection of this basic aprioristic approach to the study of science, and particularly a rejection of the imposition of canons of how science *ought* to proceed from sources outside of the sciences themselves. Ironically for our purposes, this movement in philosophy of science is sometimes called "naturalistic" philosophy of science (cf. Callebaut 1993). The word 'naturalism', of course, means something totally different here from what it means in philosophy of mind. In philosophy of science, it signifies simply the view that the sciences of nature are not beholden to standards imposed from outside of the sciences of nature themselves (standards such as making scientific explanation look like mathematical deduction, for example).

The irony here is that what naturalistic philosophy of science is rejecting is precisely the kind of position taken by those naturalistic philosophers of mind who take the normative stance that reducibility serves as a kind of litmus for the legitimacy of the special sciences. Naturalistic philosophers of *science* view philosophy of science as being itself a kind of second-order empirical inquiry, relatively continuous with the sciences themselves. This kind of inquiry is compatible with the discovery that particular *positive* claims about the nature of explanation are true (or that they are false). However, it is quite antithetical to taking the kind of normative stance toward scientific explanation found among the Positivists. This does not, of course, mean that the reductionist philosopher of mind is necessarily wrong to take such a normative stand; it could be that it is the naturalistic philosopher of science who is in the wrong here. But if the reductionist philosopher of mind wishes to make normative pronouncements about what forms scientific explanation *must* take, she will find little support in this enterprise from contemporary philosophers of science. Naturalistic philosophy of science and normatively naturalist philosophy of mind make for poor bedfellows.

3.2. Taking Real Science Seriously

The 1960s and 1970s saw an increasing trend of bringing the philosophy of science into closer contact both with the history of science and with case studies of actual scientific explanation drawn from a variety of sciences. This period saw a number of criticisms of most of the central Positivist doctrines, both from within the Positivist/Logical Empiricist tradition itself, and from a new Historicist movement that received much of its momentum from the successful reception of Thomas Kuhn's (1962) *The Structure of Scientific Revolutions*.

3.2.1. Rejecting the Derivational Model of Explanation

One of the core features of all versions of the historically influential forms of reductionism is that they treat explanation as some sort of derivation, be it on the model of syllogism, mathematical proof, axiomatic systems, or construction. The Positivist/Logical Empiricist project might indeed be viewed in part as a grand attempt to reconstruct scientific explanations in axiomatic form. The only problem is that *the attempt failed*. It was most successful in the reconstruction of mechanics (most notably, E. Nagel 1961). But the concentration on mechanics tended to obscure the fact that mechanics was, if anything, a special case in its susceptibility to axiomatic reconstruction. As Stephen Toulmin (1974, 610) writes,

> In mechanics—and in mechanics alone—the intellectual content of an entire physical science could apparently be expounded as a single mathematical calculus. Here was a complete natural science free of logical gaps and incoherences. . . . The temptation to hold theoretical mechanics up as a mirror to other branches of science, and to demand that other sciences be construed on the same model and achieve the same logical coherence, seemed irresistible. Yet the very formal perfection of the theory ought surely to have ruled it out as the "type example" of a natural science, and prevented us from extrapolating conclusions about the "logical structure" of mechanics, so as to apply to natural sciences generally. Rather, we need to recognize how *exceptional* a science mechanics really is.

Patrick Suppes (1974, 66), who had himself undertaken the project of axiomatizing a number of areas of science, likewise came to the conclusion that, while such reconstructions were possible in some areas of the sciences, they were by no means the rule.

Other writers of the 1970s and 1980s, such as Schaffner (1967, 1974), and P. S. Churchland (1986, chapter 9), argued that, even when derivational reductions are to be had, they tend to be produced well after the real scientific progress had been made in *non*derivational explanations, and that the derivational reductions were generally peripheral to the explanatory force of the science. Writers of the 1990s, such as Nersessian (1992) and Trumpler (1997), argued that derivational reconstructions of theories often misdescribed theories "in the wild" (Craver

2002), whether because the theories were captured only incompletely by the axiomatization (Craver) or because the reconstruction obscures the rich diversity of representational tools actually deployed in theories and models (Nersessian).

Perhaps most damningly, the derivational account of explanation was incompatible with known forms of explanation in the sciences. In the 1980s, philosophers of biology made this point with respect to their own science (Bechtel 1983; Wimsatt 1974, 1976a,b, 1980b; Campbell 1974; Brandon 1985; Sarkar 1992; Gould and Lewontin 1979), but the point had already been made in a more general way by Wesley Salmon (1971, 1984), particularly with regard to statistical explanation. Salmon's writings, perhaps more than any others, persuaded philosophers of science that the deductive-nomological (D-N) model of explanation needed to be abandoned, and that a successor account needed to be found. (And the criticisms of the D-N model could be extended to other derivational models as well.)

3.2.2. Scarcity of True Reductions

In retrospect, Toulmin and Suppes may in fact have been overly *optimistic* about the prospects of axiomatic reconstructions of scientific explanations in the early 1970s. In fact, even the most frequently cited examples of successful intertheoretic reductions have been shown to be "gappy." Given that it is this very core of examples that are used again and again as intuition pumps for some form of reductionism, the failures of reduction in even these core cases cannot help but be significant. As Michael Silberstein (2002, 94) writes in *The Blackwell Companion to Philosophy of Science*:

> Focus on actual scientific practice suggests that either there really are not many cases of successful epistemological (intertheoretic) reduction or that most philosophical accounts of reduction bear little relevance to the way reduction in science actually works. Most working scientists would probably opt for the latter claim. Often discussed cases of failed or incomplete intertheoretic reduction in the literature include:
>
> - the reduction of thermodynamics to statistical mechanics (Primas 1991, 1998; Sklar 1999)
> - the reduction of thermodynamics/statistical mechanics to quantum mechanics (Hellman 1999)
> - the reduction of chemistry to quantum mechanics (Cartwright 1997; Primas 1983)
> - the reduction of classical mechanics to quantum mechanics (such as the worry that quantum mechanics cannot recover classical chaos) (Belot and Earman 1997)

To this list one might well add the arguments in philosophy of biology that molecular genetics cannot provide a derivation base for evolutionary biology (Lewontin 1983; Levins 1968) or even for classical genetics (Kitcher 1984).

A detailed exposition of each of these failures of true reduction is beyond the scope of this small section; but perhaps it is worth a brief detour through discussions of what is perhaps the most cited example of a putatively successful reduction: the derivation of thermodynamic results (or, more narrowly, of the Boyle-Charles gas law) from statistical mechanics. One problem with Nagel's attempt to perform this derivation, pointed out by Alan Garfinkel (1981/1999), is that E. Nagel's derivation requires ancillary assumptions (i.e., assumptions not present in statistical mechanics itself)—assumptions which, moreover, turn out to be false. In particular, Nagel (1961, 344) writes:

> A further assumption must be introduced . . . that the probability of a molecule's occupying an assigned phase cell is the same for all molecules and is equal to the probability of a molecule's occupying any other phase cell and (subject to certain qualifications involving among other things the total energy of the system) the probability that one molecule occupies a phase cell is independent of the occupation of that cell by any other molecule.

Garfinkel argues that this assumption is incompatible with two standard assumptions: the conservation of energy and the normal distribution of velocities (456). From this, Garfinkel concludes:

> The failure of these independence assumptions tells us that we do not really have a case of a global property arising as a simple aggregate of independent individuals. There is, to be sure, a collection of individuals (the gas molecules) with an individual nature given by Newtonian mechanics, according to which they are essentially small elastic particles. But the properties of the gas, like the Boyle-Charles law, do not arise simply from this individual nature. We must make, in addition, strong assumptions about the *collective* possibilities of the system, assumptions which are imposed on the individual nature and do not in any sense follow from it. Their effect is like the effect of the kinematical conditions discussed earlier [in Garfinkel's article]: to restrict sharply the a priori possibilities of the system. (456–57)

The upshot of this is not, of course, that statistical mechanics explains nothing about thermodynamics in general, or about the Boyle-Charles law in particular. Rather, the point is that it is not a conceptually adequate microexplanation without remainder—that is, it is not a broad reduction.[1]

Problems with reduction have stood out even more strikingly in philosophy of biology. Initially, after the discovery of DNA, there was a trend in the life sciences to regard *molecular* biology as the most important level of investigation, and to turn attention away from historical processes of evolution and ecological relations between organism and environment. Molecular biology was widely assumed to provide a reduction of classical biology, or at least of classical genetics, as well as to afford the opportunities to study physical mechanisms underlying both inheritance and development. However, there

has been an important backlash against this assumption in evolutionary biology, as expressed in this excerpt by Harvard biologist Richard Lewontin (quoted in Callebaut 1993, 261):

> Any textbook or popular lecture on genetics will say: "The gene is a self-reproducing unit that determines a particular trait in an organism." That description of genes as self-reproducing units which determine the organism contains two fundamental biological untruths: *The gene is not self-replicating* and *it does not determine anything*. I heard an eminent biologist at an important meeting of evolutionists say that if he had a large enough computer and could put the DNA sequence of an organism into the computer, the computer could "compute" the organism. Now that simply is not true. Organisms don't even compute themselves from their own DNA. The organism is the consequence of the unique interaction between what it has inherited and the environment in which it is developing (cf. Changeux 1985; Edelman 1988a, b), which is even more complex because the environment is itself changed in the consequence of the development of the organism.

There are really two antireductionist strands here. First, biologists like Lewontin and Stephen Jay Gould (Lewontin and Gould 1979; Lewontin 1983) and philosophers like Philip Kitcher (1984) have argued that it is a misconception to view the relation between molecular biology and classical genetics or evolutionary theory as a reduction. First, while genes are a causal factor in determining phenotypic traits through development, they do not "determine" those traits. Every trait is a product of a combination of genetic inheritance and environmental influence through development. Phenotype is "plastic," and sometimes quite dramatically so. Second, both the theory of evolution and developmental biology require us to view living organisms *historically* (over evolutionary and developmental time frames, respectively) and as systems that are open to their environments. Key biological notions like "fitness" are inherently relational: the fitness of an individual organism is determined *jointly* by its internal traits and its environment, and a trait that is adaptive in one environment would be maladaptive, perhaps fatally so, in another. Because fitness is a key element of the historical story that must be invoked to explain an evolved trait, molecular genetics (even supplemented by developmental biology) leaves out the resources needed to explain why a present-day population has the array of genes it does. This requires a kind of explanation that appeals to larger systems (organisms and populations in particular environments over time) to explain smaller ones (the existence of particular genes in a present-day population). Moreover, biology requires a number of types of interlevel causal stories to accommodate other machinery of evolutionary biology, such as coevolution, exaptation, niche selection, and niche construction. One must move back and forth between levels of the gene, mechanisms for gene expression in development, individual organisms, populations, and

environments, often employing "downward" explanation of smaller structures in terms of the history and dynamics of larger systems. Worse still, key concepts in these different levels, such as "organism" and "environment," are so thoroughly tied to one another that they are essentially interdefined.

While such insights first came to light in connection with evolutionary biology, they have turned out to be applicable to other scientific explanations as well, in such fields as the biochemistries of metabolism or fermentation, and neuroscience (cf. Bechtel and Richardson 1993. We shall see in chapter 7 that such features are sometimes present in physical explanations as well). One main locus of such work has come in the study of "mechanistic explanations."

The fact that even the supposed paradigm cases of intertheoretic deductions do not, upon scrutiny, turn out to fall into the mold of broadly reductive explanation probably does little to weaken the explanatory insights to be found at the interstices between sciences. But it does give the lie to a particular *metatheoretical* perspective on these, as represented by several of the key positive explanatory (EP) theses of broadly reductive naturalism:

EP 1: Broadly reductive explanations have played an important role in the natural sciences.
EP 2: There are successful interdomain "reductions" by way of broadly reductive explanation.
EP 3: Broadly reductive explanation is (in fact) the only form of explanation employed in the natural sciences.

Even the comparatively weak EP 2 seems now to be threatened. And even if we can find a few genuine examples of broadly reductive explanations, these probably still do not amount to enough to justify EP 1, much less EP 3. And without these there seems to be little justification for believing the further claims EP 4 and EP 6:

EP 4: There is a "basic" science to which all other "special" sciences can be reduced by broadly reductive explanation, that is, such that all more complex phenomena can be constructed from it (at least in principle).
EP 6: Given a complete physical description of the world at time T, one could (in principle) derive *all* facts about the world at T ("vertical" prediction).

Only aprioristic reasons would then remain for holding the normative claims about explanation (EN) EN 1 and EN 2:

EN 1: Broadly reductive explanation is the only *legitimate* form of scientific explanation.
EN 2: For any special science S, either S can be reduced to basic physics through broadly reductive explanation, or else the legitimacy of S is undercut (and with it our warrant for believing in the objects postulated in S).

In short, explanation is not generally derivation, and indeed the majority of scientific explanations cannot even be successfully *reconstructed* as reductions.

Without this assumption, the rest of the metatheory seems to be left without a foundation.

3.2.3. Alternative Metatheories of Explanation

No single metatheory of explanation now enjoys the consensus status that was enjoyed by the deductive-nomological model in the 1950s. Among the leading contenders are more relaxed forms of empiricism, which reject some of the characteristic assumptions of Positivists and Logical Empiricists (such as verificationism and the modeling of explanation upon the logical syllogism), but hold onto other empiricist themes (such as a Newtonian emphasis on the central role of laws); *causal* accounts of explanation, which hold that the only thing that can give an explanation of a singular event is its causal etiology; *mechanistic* explanation; and *pragmatic* or *erotetic* accounts, which view explanations as answers to very particular *why*-questions. (These accounts need not be viewed as mutually incompatible. Philosophers of science such as Bas van Fraassen [1980] and Philip Kitcher [1981] each combine empiricist and Pragmatist elements in varying degrees,[2] while Nancy Cartwright [1989, 1999] arguably combines Pragmatist and causal themes with some amount of residual empiricism.[3])

About causal accounts, such as that of Salmon (1984) or Lewis (1986), I shall say little here, except to signal that they are most compelling in the case of singular explanation, that is, explanation of particular events and not of intertheoretic relations. Moreover, neither causal nor nomic accounts would, if applied to the mind, yield anything like mind-body reductions. Laws or causal relations, whether relating mental states to one another or to brain states, are compatible with a variety of metaphysical interpretations. And causal relations, at least, absolutely require that relations be between distinct objects and/or events.

Mechanistic explanation is in some ways an heir to the reductionist tradition, as it is concerned with the examination of the mechanisms through which a process—say, metabolism—is achieved, and proceeds through decomposition of a larger system into subsystems and parts. However, careful scrutiny of how such mechanisms are modeled in real science often reveals that they are badly described by the reductionist model. Such explanations often require causal links *between* distinct models. Often, the systems thus related do not stand in a part-whole relationship to one another. But even when they do so, the structure of the more inclusive system can act as a "control hierarchy" that influences the behavior of the parts:

> In a control hierarchy the upper level exerts a specific, dynamic constraint on the details of the motion at the lower level, so that the fast dynamics of the lower level cannot simply be averaged out. The collection of subunits that forms the upper level in a structural hierarchy now also acts as a constraint on the motions of selected individual subunits. This amounts to a feedback path between levels. Therefore, the physical behavior of a control hierarchy must take into account at least two levels at one time. (Pattee 1973, 77)

Mechanistic explanations of the sort described by Pattee and by Bechtel and Richardson share with reductions a decomposition of a system into subunits that divide the labor, but they often violate the reductionist assumption that interlevel explanations must only proceed "upward" from parts to wholes.

The pragmatic or erotetic account of explanation strikes me, by contrast, as standing in opposition to the received view of the seventeenth and twentieth centuries in a much more fundamental way. Even though Early Modern science was virtually founded on the realization that some of our ways of apprehending the world (particularly through the senses) do not reflect its true nature, both Early Modern Rationalist and twentieth-century Positivist/Logical Empiricist approaches to science involved a tacit assumption that our scientific language could reflect real and even essential properties of the material world in a fashion that would require no hedging by references to the pragmatic contexts in which that language is used and in which explanations are offered. As a result, explanation could be cashed out purely in terms of semantic and syntactic relationships between statements, that is, as a kind of derivation.[4]

The pragmatic account, by contrast, locates the sine qua non of explanation within the enterprise of asking and answering questions about why or how things behave as they do. Once one has taken this turn, the enterprise of explanation looks very different. To assess what could count as an explanation, one must first locate the question to which the explanation would provide an answer, and then one must take into account what information is *missing* in such a fashion as to give rise to the question in the first place. Once we have made this turn, several things seem to fall into place all at once:

1. The status of an explanation as an explanation cannot be read off its syntactic and semantic properties alone.
2. The semantic properties of a sentence that are relevant to its being an explanation cannot be read off its grammatical form alone, as these depend on its relation to a larger pragmatic enterprise.
3. As there are many kinds of questions, there are likely to be correspondingly many different types of explanation, and these are unlikely to be typed on merely syntactic or even semantic grounds.
4. The status of a given claim, model, or theory as an explanation would seem to require a second and broader model of the pragmatic context, and not merely of the phenomena modeled; for example, a model of the cognitive and/or social activities of which questioning and answering are a part, what are taken as background assumptions, acceptable margins of error, and so on.

These observations strike me as congenial and even obvious, at least so long as pragmatic accounts of explanation are not made into a new kind of reduction of science to nothing more than its status as a set of cognitive or social activities. But Pragmatists in philosophy of science are generally *not* inclined to make the objectionable sort of move that robs the world of its role in making theories *true* or reduces the notion of "truth" to what one's colleagues will allow one to get

away with saying. Rather, the view is that pragmatics plays a role in cashing out the nature of the questions with which we interrogate the world and the space of possible answers that goes along with the questions, but the world itself plays a decisive role in (a) determining which answers are *true*, and indeed (b) which sorts of questions are *apt* ones. (We now know, for example, that questions framed in Ptolemaic terms are not apt questions.)

My own preferred take on the pragmatic turn, which is explored in greater depth in the final part of the book, emphasizes the cognitive rather than the social elements. On the cognitivist view, the sine qua non of explanation is its role in producing a gain in understanding. Such a view would be filled out in greater detail by a theory of the kinds of understanding that actually take place as a result of scientific explanations, and perhaps of explanations more generally. Such a theory might emphasize such activities as the formation, testing, and refinement of *mental models* of features of the world, the explanation of particular events by the deployment of such models, and the various ways that separate models can be brought into contact with one another. This latter activity might include theory reductions when they are to be had, but might include much more partial and incomplete relations as well. In short, it is not at all clear that a Pragmatist/erotetic account of explanation will offer much hope of the revival of the project of finding a reductively unified science. Indeed, it arguably leads in precisely the opposite direction: the different questions that lie behind different explanations create barriers to their integration. These implications of the Pragmatist approach to explanation also seem consonant with two other important thematic developments in Post-Empiricist philosophy of science: the plurality of explanatory types and the (limited) autonomy of local scientific domains.

3.2.4. The Plurality of Explanatory Types

One of the first elements of the Unity of Science program to be rejected was the view that the various sciences should all practice the same methods of inquiry and explanation. This idea had in fact already been rejected by Oppenheim and Putnam in their 1958 "Unity of Science as a Working Hypothesis." And their rejection of this doctrine was all the more confirmed by subsequent case studies of current and historical explanatory practices in the sciences. Two cases already mentioned—statistical explanation and evolutionary explanation—should perhaps suffice to make the point that different sciences employ, *and require*, different explanatory strategies. This does not, of course, imply that just anything can count as a good explanation, or that the sociological fact that scientists at a given time count something as a successful explanation is enough to make it so. The claim is more modest, but still quite significant: namely, that (a) our metatheory of explanation needs to be beholden to scientific practice and not (just) to aprioristic philosophical standards, and (b) the sciences in fact employ a variety of types of explanation, and (c) what *makes* a given explanatory type apt or even obligatory is determined in large measure by features internal to a specific scientific domain.

3.2.5. Autonomy of Local Domains

This emphasis on the relation of explanatory virtues to the aims of a particular science can also be viewed as a claim for the *methodological autonomy* of the special sciences. This and other claims about the autonomy of the special sciences became an important theme among historicist philosophers of science in the 1960s and 1970s. Historicists like Larry Laudan (1977), Dudley Shapere (consolidated in Shapere 1984), and Stephen Toulmin (1972) attempted to rethink issues in philosophy of science out of case studies. Joseph Rouse (1998, 73–74) characterizes the important features of this movement as follows:

> The most widely noted lesson these postpositivist historicists drew from the failure of positivism was the importance of attending to the details of particular sciences. They took the positivist tradition to have developed theories of confirmation and explanation that were inconsistent with the actual historical development of the sciences. Their response was to insist that scientific methodology was "domain-specific" and/or interdependent with a field's theoretical commitments. Along with their commitment to historical specificity came a rejection of formal methods in the philosophy of science: the logical structures of induction and probability, the paradoxes of confirmation, the "Received View" of theories as axiomatic calculi that invited semantic reductions of theoretical vocabulary, and the formal problems arising within deductive-nomological accounts of explanation, all exemplified the positivists' supposed failure to connect with scientific knowledge. Philosophy of science should be determined by the historical development of science, not by prior philosophical commitments in epistemology or semantics.

Two features of this historicist turn seem particularly important for our purposes: the domain-specificity of scientific methodology and the central importance of progress *within* a scientific domain in the maturation of that science. Indeed, the historicist turn forces us away from a global notion of "Science" and toward a more pluralistic notion of science*s*, each of which may have features that are significantly different from those of other sciences, and is largely justified by its own internal successes.

This theme, sometimes called the *autonomy of the special sciences*, is of signal importance for our investigation. For it seems to involve a repudiation of the normative claim EN 2 (the claim that the special sciences must be "vindicated" by a demonstration that they are derivable from more basic sciences) as a *general* methodological canon in philosophy of science. If EN 2 is not a general methodological canon, it is unclear why it should have any *special* appeal when the science in question happens to be psychology. There have been a few publications in philosophy of mind that have turned this corner, notably Stich and Laurence (1994) and Baker (1995), both of which point out that

the "reduce-or-eliminate" standard could not be applied to *other* special sciences, or indeed to many commonsense domains, without depriving us of an untold number of explanations that have not otherwise been called into question. But, by and large, philosophers of mind have not yet caught on to this lesson. If there is in general a forced choice between reduction and elimination, then we would have to eliminate most of the ontological inventories of the special sciences, and this consequence is generally seen as intolerable. But if domain autonomy is the general principle in the sciences, then the application of a reduce-or-eliminate principle in the special case of psychology would require special pleading, and not the kind of appeal to general scientific principles generally offered by eliminativists.

3.2.6. Plurality of "Good-making" Qualities

This recognition of the differences between individual scientific disciplines also called for a rethinking of what makes for "good science." The demarcation problem—of finding a single criterion that distinguished scientific from nonscientific and proto-scientific enterprises—having been abandoned, the door was open for a more ecumenical discussion of what Newton-Smith (1981) calls the "good-making qualities" of scientific theories, models, and explanations. Even a defender of scientific rationality like Newton-Smith is inclined to posit that there can be a variety of separate good-making qualities or explanatory virtues, each of which contributes to the epistemic quality of scientific understanding. None of these may be individually necessary for the status of science, and different sciences may enjoy some or all of them in different degrees. A number of philosophers of science have offered lists of such good-making qualities (e.g., Newton-Smith 1981; Thagard 1988, 1992). However, one might also be disposed to think that this list is essentially open-ended, as there may be additional virtues that are most relevant to the scientific disciplines of the next century, or indeed to existing scientific disciplines that have not been adequately studied.

3.2.7. Variety of Interdomain Relations

Finally, the rejection of reductive models of relations between scientific domains has been accompanied by more careful study of the kinds of relations that really have been found at the interstices between particular sciences. The rejection of the paradigm of intertheoretic reduction and the acknowledgment of the justificatory and methodological autonomy of individual scientific domains ought not lead to the conclusion that the various sciences are each free-spinning wheels, unconnected to one another. Rather, it opens the door to investigation of what kinds of relations are actually to be found through a historically faithful examination of cross-disciplinary work.

The most influential metatheoretical notion that has been proposed in this area is Darden and Maull's (1977) notion of *interfield theories*. Darden and Maull develop this idea through examining a number of cases in which researchers

have integrated ideas from (what were at the time) two (or more) disciplines. In one such example, Darden and Maull explore how insights from cytology (study of cells) were linked to insights from Mendelian genetics by Boveri (1903) and Sutton (1903). Before 1903, cytology and the study of heredity had constituted separate fields. The discovery of the chromosome initially arose as a physiological notion in cytology. Chromosomes were identified as a particular kind of body observed in the cell nucleus. At the same time, the Mendelian model of heredity was employed without a theory of the physical basis of inheritance. Boveri and Sutton suggested that the units of Mendelian heredity were located in the cytologist's chromosomes. This identification was to become the basis of classical genetics of the Morgan school.

In this case, one discipline (cytology) provided materials needed to answer problems that arose in another (Mendelian genetics), and the result was the birth of a new kind of theory (classical genetics) that spanned both preexisting fields. These features—the turning to a field A for resources to explain an existing problematic in a field B, and a new theory that spans parts of both A and B—are typical of the cases examined by Darden and Maull. However, it is not clear that either of these features is necessary for there to be important relations that cross the boundaries of disciplines and domains. Bechtel (1984), for example, offers a case study of the links between vitamin research and metabolism research (specifically, the role of B vitamins in respiratory coenzymes). Here the interfield connection was not arrived at as a result of pursuing a preexisting problematic in either discipline, and required a reconceptualization within each of the fields. Likewise, one discipline may borrow from another without the creation of an entire new discipline, or even take the form of a theory that is best viewed as spanning previous disciplinary boundaries.

The expression ''interfield theory'' seems in fact to be doing double duty in philosophy of science. It is used *broadly* as a rubric for any discussion of relations found at the boundaries of two scientific disciplines; but it is also used *narrowly* for the particular *kind* of relations characteristic of Darden and Maull's case studies, such as the formation of a theory (or even a new discipline) that straddles or redraws previous disciplinary boundaries. The conflation of these two usages strikes me as having effects contrary to Darden and Maull's deeper aim of doing justice to the variety of relations one finds at the boundaries between disciplines. As a result, I should prefer to restrict the expression 'interfield theory' to the narrow use, and employ a more ecumenical generic term, such as *connective explanatory virtues*, as a collective term for any of a variety of connections between domains that help to confer understanding. One would then name such connections individually: for example, *explanatory borrowing*, in which insights from one domain are employed piecemeal within explanations in another domain, or *identification of elements*, in which items from two domains (e.g., chromosomes and genes) are hypothesized to be identical, but *without* the kind of wholesale theoretical integration required by broadly reductive explanation or CAEs.

One line of investigation in interdisciplinary relations that I find particularly promising makes use of Herbert Simon's (1977) taxonomy of types of

systems, and particularly his distinction between those that are "decomposable" and those that are not. This line of research, carried out by Simon and later by Bechtel and Richardson (1993), promises a framework for classifying kinds of relations between their systems and their parts which would additionally provide criteria for *which* sorts of systems are susceptible to broadly reductive explanation.

The overall point here is that there *are* in fact a variety of fruitful ways that two scientific domains can come into contact with one another, but that fall far short of the kind of derivation relation that is distinctive of broad reduction. This is a *general* point in philosophy of science. But this general point can also be applied to case studies in psychology and neuroscience as well, for example, the relations between psychophysics and localization. In this respect, the partial connections between different enterprises that attempt to understand cognition do not seem out of step with what one finds in the natural sciences after all—*not* because they involve a reduction to the natural sciences, but because the kinds of nonreductive relations to be found there are broadly continuous with the kinds of interdomain relations one finds among the sciences generally. But to appreciate this variety of interdomain relations, one must leave the reductionist metatheory of the mid-twentieth century behind and embrace an approach to philosophy of science that seeks to understand the various forms of explanation one actually finds among the sciences, and that respects both the autonomy of local domains and the forms of partial explanation found at their intersections.

3.3. Scientific Disunity

As the Positivist project of unifying the sciences was officially known as the "unity of science" movement, the themes discussed in this chapter are often lumped together under the heading "disunity of science." Likewise, philosophers of science are today increasingly disinclined to speak of "Science" as a generic singular term—especially in a way that even implicitly involves capitalization—and to speak of "sciences" in the plural.

It is important to stress here that *none* of these themes is any way "anti-science." Despite the claims of well-known figures like E. O. Wilson (1998) in *Conciliance* and Francis Crick (1993) in *The Astonishing Hypothesis*, the most important criticisms of reductionism have come, not from Postmodern relativists or fundamentalist religious Luddites, but from historians and philosophers of science. The issues here are *not* about the accomplishments, laudability, or moral fiber of the sciences or of scientists, but of the right metatheoretical characterization to give to explanations within scientific disciplines, and explanations that span the boundaries of two or more disciplines. Enlightenment Rationalists and Logical Positivists favored a reductive metatheory, but did so largely on armchair, aprioristic grounds. To the extent that one has reason to trust armchair reasoning to lay down norms for the shape of the sciences, one *might* even still be inclined to view this as a tenable

normative project. But to the extent that the philosophy of science is guided by careful examination of how real science is done, this metatheoretical picture does not seem to stand up to much scrutiny. And, in my opinion, any attempt to resurrect reductionism as a normative metatheory for the sciences would need to proceed by way of equally careful case studies in order to merit a second consideration. (Some philosophers, such as John Bickle [1998, 2003] have attempted to rejuvenate things called "reductionism," but Bickle's "reductions" involve only contingent identifications of elements from different domains, and he agrees that this is weaker than classical notions of reduction and is not enough to underwrite metaphysical supervenience.)

There are, of course, a variety of positions within the "disunity of science" camp, as that label itself is in large measure simply a dismissal of the Positivist unification program. At the most conservative end of the spectrum one might find people like Philip Kitcher (1981), who rejects reductionism, and yet does so in part by comparing it unfavorably with another kind of connection which he calls "unification," consisting in uniting phenomena under more general and powerful categories. Kitcher's unifications seem closer to the Newtonian than the Galilean model of explanation. (And indeed, Newton's mechanics, which unified the ballistics with celestial mechanics, is perhaps the example par excellence of "unification" in Kitcher's sense.) Yet even Kitcher's unificationism is offered in the spirit of saying that we should *look for* such unifications, and that they are a good thing when we can find them. (Who could object to that?) Kitcher does not take any sort of global unification to stand as a norm for the legitimacy of the special sciences. A similar view is expressed by Ian Hacking (1996), who contrasts the notion of "unity as singleness"—for example, singleness of method, or singleness in the form of a comprehensive deductive system spanning multiple scientific disciplines—with "unity as harmonious integration" of work in separate scientific domains. He thinks the former unlikely, but holds out some hope for the latter. At the more radical end of the spectrum is John Dupré (1993), who views the disunity of the sciences as an indication of a deeper ontological disunity: the sciences are radically disunified because the world is radically disunified. Perhaps somewhere in between one might find Nancy Cartwright (1999), who characterizes the relations between the sciences as a kind of patchwork, portraying a "dappled world."

3.4. Implications for the Mind

The majority of broad reductionism's claims about the explanation of mental phenomena are bound up with a particular metatheory of the nature of explanation generally in the sciences, and about the nature and status of interdomain explanations in particular. Its empirical hypotheses about the mind are guided by the idea that mental phenomena might turn out to fit into this metatheoretical framework; its normative claims are likewise guided by the idea that this metatheoretical framework holds normative force for the special sciences. The commitment to the metatheory, in turn, is generally

based on the supposition that it has been established on inductive grounds as a metatheoretical picture for the sciences generally. Out of this comes the conclusion that psychology and the mental stand in need of "vindication" by way of demonstrating their compatibility with this larger picture.

The problem I have developed in this chapter is that the general metatheoretical picture in question has largely been rejected by philosophers of science, and thus philosophy of psychology at the turn of the millennium is still holding itself hostage to the demands of philosophy of science of the 1950s. The special sciences do not, in general, require vindication by intertheoretic reduction, and there is no special reason to impose such a standard when the science in question is psychology (compare Stich and Laurence 1994). Moreover, far from being the norm in the natural sciences, broad reductions are in fact quite rare, usually require some philosophical reconstruction, and the reconstructions that permit axiomatic formulation are only peripherally relevant to the explanatory force of the original theories. If there are explanatory gaps between psychology and neuroscience, there is no obvious reason to see these as more threatening to the status of psychology than explanatory gaps between chemistry and biology are to the theory of evolution. Moreover, there are a variety of types of explanation at work, both within particular scientific disciplines and at the borders between disciplines. If psychology employs some distinctive and proprietary forms of explanation, such as intentional and rational explanation, this fact in itself need not constitute a *problem* for psychology, any more than the distinctiveness of, say, explanation by natural selection constitutes a problem for biology.

In short, reductionism's anticipations of how psychology might be accommodated within the natural sciences, and its problematic for the philosophy of psychology, are predicated on assumptions about explanation generally that have been rejected—and rejected for convincing reasons—by philosophers of science. Insofar as we want our understanding of the mind to be informed by the best understanding of the sciences available today, we need to move *beyond* the reductionist view of science to do so.

What implications does this have for philosophy of mind? One implication that seems clear-cut is that these developments are bad news for reductionism and the forms of eliminativism that are the flip side of reductionism. The appearance that there were abiding explanatory gaps in psychology was bad enough for the reductionist. But so long as one believed the possibility of intertheoretic reductions to be well-established by a long catalogue of reductive successes in the natural sciences, there was at least a zealot's hope that the appearance of psychological gaps would prove to be illusory. One might even have taken widespread reductions to be reason to hold to reductionism as a normative claim. But if it is "gaps all the way down," as it were, such a hope seems foolhardy. If there are explanatory gaps within physics itself, or between physics and chemistry, or molecular and classical genetics, finding reductions for consciousness or intentionality would be quite remarkable. Indeed, it would mean finding a *stronger* link between mind and brain than is generally found between two domains of the natural sciences.

At first glance, this is good news for friends of the explanatory gap. Finding that explanatory gaps are commonplace helps to undercut any suspicion that might have accrued to a unique gap between mind and brain. And this is helpful both to dualists and to nonreductive materialists. But what is given with one hand is taken away with the other. If we conceive of a natural world united by explanatory reductions (and hence via the Positive EMC Principle by metaphysical supervenience), the occurrence of unique explanatory gaps in the case of the mind presents a remarkable, fascinating, and sexy problem. But if the psychological gap is just one gap among many, the problem no longer seems so remarkable, fascinating, or sexy. And if friends of the gap combine it with the Negative EMC to argue that mental phenomena do not supervene upon physical facts, their position is now complicated by the fact that such an inference should equally entail that chemistry and biology do not supervene upon physics. Even nonreductive physicalists like Davidson have often assumed the mind to be *unique* in its irreducibility. And Mysterians like McGinn and Nagel have characteristically located the source of the mystery in some feature special to cases where the mind studies itself.

This problem will prove particularly acute for dualists. The dualist will face an uncomfortable dilemma. If he continues to embrace negative EMC, he will be pushed to count beyond two and become a pluralist of a higher ordinality. But if he blocks this move toward pluralism by renouncing Negative EMC, he deprives himself of the principal argument traditionally used to argue in favor of dualism. In the face of scientific pluralism, dualism *with* Negative EMC would seem to be inconsistent; dualism *without* Negative EMC would appear to be consistent, but without much residual argumentative support.

Nonreductive physicalism faces a similar problem. Most nonreductive physicalists reject Negative EMC, and hence are not threatened by the pull of radical *ontological* pluralism on grounds that derive it from scientific theory-pluralism. Nor need nonreductive physicalists be threatened by the rejection of intertheoretic reduction in the sciences. While nonreductive physicalists have often assumed that nonmental properties are reducible to basic physics, rejecting this assumption is quite compatible with nonreductive physicalism. It just leads to a nonreductionism *all the way down*. So far, so good. But historically, the principal reasons for thinking that physicalism might be *true* were found in reasons to think that the phenomena of the special sciences could be reductively *explained*. So long as one assumed that this was true for the natural sciences, it seemed like a reasonable inductive hypothesis that physical facts determine all the facts, even if there are isolated corners of the world where we cannot understand how this might be so. But in light of scientific pluralism, the nonreductionist must return to the drawing board, and ask anew what the best metaphysical interpretation of science is, *given the fact of widespread explanatory gaps*. From this standpoint, it is not clear that there is any reason to prefer a physicalist interpretation over its alternatives. Like dualism without Negative EMC, nonreductive physicalism is *consistent*, but it is not clear why we should believe it to be *true*.

However you slice it, someone has a lot of explaining to do. In fact, just about everybody has some explaining to do.

This is, in some ways, philosophical *terra incognita*. And so the first thing that is in order is to try to survey the landscape by considering the possible harbors that are available in this new world. The chapters comprising part II therefore examine the prospects for reductive and nonreductive physicalism, dualism, and eliminativism at greater length, and make a case that scientific pluralism gives us reason to look for a very different sort of theory.

PART II

Philosophy of Mind and Post-Reductionist Philosophy of Science

This page intentionally left blank

4

Reductionism and Eliminativism Reconsidered

The most obvious implication of theory pluralism for philosophy of mind is that it spells bad news—perhaps decisively bad news—for reductionism as a thesis about explanation and for reductive physicalism as a metaphysical thesis. Indeed, one might well be inclined to see this case as open and shut and wonder why reductionism and reductive physicalism have not disappeared from the philosophical landscape already. But old views die hard, and so I will risk belaboring the point in an attempt to salt the earth so that nothing grows on this particular soil ever again. I shall begin by recapitulating some of the principal attractions of reductionist views, then spell out the problems for broadly reductive explanation of the mind, and finally explore the connections between reductive explanation and its corresponding form of physicalism. Along the way, I shall also develop problems for at least one type of eliminativism, the type that is basically the flip side of reductionism in holding that there is a forced choice between reduction and elimination, and hence that the only alternative to reduction is elimination. In the face of theory pluralism, this view would commit us to the elimination of most or all of the entities of the special sciences.

4.1. Reduction's Seductions

Philosophers and scientists alike have been attracted to reductionism on a number of grounds. Much of its origins, in both the seventeenth and twentieth centuries, arose out of attempts to confer upon scientific explanation the same degree of deductive rigor found in

mathematics (a motivation that we might, in retrospect, be inclined to view as "math envy"). For the Early Moderns, such a project was almost forced upon them by the then-operative[I] notion of *scientia*, which was modeled on logical and mathematical inference. This led the Rationalists, in particular, to the perhaps absurdly unrealistic notion that one could even know physical first principles *indubitably* through reason alone. But even Empiricists like Locke and Hume were ambivalent about calling empirical science "knowledge" (*scientia*) because of their respective skeptical worries. (In spite of their deference to what they conceived as "Newtonianism," Newton himself was arguably a better philosopher of science, and was far more eclectic and practical in his epistemology.)[2] In the twentieth century, "math envy" also had a certain influence, as the Positivists attempted to do for the sciences what they viewed Whitehead and Russell's *Principia Mathematica* as having done for arithmetic: that is, to show that the phenomena of the special sciences could be seen as constructions out of either basic physical objects or sense data. In the case of psychology, this was perhaps most clearly seen in analytic behaviorism, which actually attempted to (re)define the psychological in terms of constructions out of stimuli and behavioral responses, and was perhaps reprised later in the century in the form of analytical functionalism.

A second motivation in the Positivist period was the project of arriving at a Unified Science. There was much disagreement among Positivist/Empiricist philosophers over what form such a unification might take. But the prevailing approach treated explanations, both within a science and of intertheoretic relations, as logical syllogisms, sometimes augmented by things like bridge laws or constructions. This was, to be sure, not the only possible vision for "unifying" science: Otto Neurath, the editor of the *Encyclopaedia of Unified Science*, preferred a model that was, appropriately enough, encyclopedic rather than axiomatic (Cat, Cartwright, and Chang 1996). The greater popularity that the reductive approach enjoyed from Carnap through Nagel is perhaps due in large measure to the overwhelming acceptance of the D-N model of explanation as a kind of rational norm for good science. It is, of course, possible to hold to a D-N model of explanation *within* a science while rejecting intertheoretic reduction as a norm. However, the deductive/axiomatic conception of explanation exerted a kind of intellectual gravity pulling in the direction of seeing intertheoretic relations as *needing* to take something like the form of reconstructions of the special sciences as derived from basic physics in an axiomatic system—the kind of project undertaken most notably by Nagel.

While there was a significant a priori element to Logical Positivist/Empiricist philosophy of science, it is not clear just how well such a philosophical interpretation of science could have fared without any supporting evidence from the progress of the sciences themselves. But many important scientific advances *do* take the form of part-whole explanations in which at least a significant set of the features of the higher-level system can be understood as consequences of, and derivable from, the features of the proper parts of the system. While these might not take the form of the contact-interactions envisioned by Early Modern mechanists, that was really unimportant: one *can*

explain valences by appeal to charged particles, the gas laws by appeal to particle collisions, and the mechanisms of biological inheritance by appeal to chromosomes and ultimately DNA molecules. Such part-whole explanations may be partial and incomplete, but they represent an important and recurring successful explanatory strategy in the sciences. And indeed, this is often all that scientists mean when they speak of ''reductions.''

Of course, part-whole explanations are not enough for what I am calling *broad reductions* unless they are also explanations *without remainder*, that is, CAEs. But if it is an important methodological principle to always *look* for comprehensive part-whole explanations, it is easy to slip into the dialectical error of assuming that they *must* be there to be found and particularly that they *must be* there to be found wherever there is currently anything that looks like an explanatory gap. Such an assumption might even be appropriate as a *working assumption* in scientific research. It is more problematic as a philosophical claim about what sorts of explanations are really there to be found. Here, I think, the combination of (a) the successes of part-whole explanatory strategies (in providing weaker and more partial forms of explanation) with (b) the normative and aprioristic Positivist views about ''the logic of science'' served to make reductionism a psychologically compelling view, even if it was by no means decisively confirmed as a second-order empirical hypothesis. After all, each of the part-whole explanations that we *have* has filled in what was once a de facto explanatory gap, and so it might not seem at all unreasonable to assume on inductive grounds that, with enough filling in, one might arrive at a unified science in the form of a grand axiomatic reconstruction based in fundamental physics.

Historically, scientists and philosophers alike tended to be divided on the question of whether such a grand reductive project could be extended to two troublesome areas: life and the mind. Even such luminaries as Newton were tempted by vitalism. However, research in cytology, molecular biology, metabolism, and developmental systems in the late nineteenth and twentieth centuries began to fill in the picture with compelling microexplanations of many features of living organisms. And as Papineau (2002) argues, this success, combined with the burgeoning of neuroscience, supplied not-unfounded hopes that similar advances would be forthcoming for many features of the mind as well. And given the *functionalist* orthodoxy of the late twentieth century, such hopes could realistically be seen as panning out, at least at the level of supplying *partial* microexplanations of a wide range of *functionally* typified psychological phenomena. Bracketing the concerns at every level about whether such explanations amounted to broad reductions, it seemed natural to hope that eventually we would see the kinds of microexplanatory successes in psychology that we had seen over four centuries in the physical and life sciences.

For philosophers, particularly metaphysicians of mind, reductionism has another and distinct appeal. Broad reductions supply a form of explanation that could potentially decisively settle long-standing metaphysical questions. If A is broadly reducible to B, then B→A is metaphysically necessary, and A is

metaphysically supervenient upon B. To the extent that it looked as though the sciences were rapidly filling in explanatory gaps by way of part-whole explanations, and might do so without remainder, this presented the exciting possibility that the sciences might decisively settle perennial philosophical debates between dualism and physicalism. If physicalism is the view that all facts about the world (and in particular, the mental facts) are metaphysically supervenient upon the physical facts, then producing reductions of mental phenomena would settle the debate decisively in favor of physicalism. And so, if reductionist claims in philosophy of mind at midcentury greatly outstripped the actual evidential base in psychology and neuroscience, the allure of reductionism can nonetheless be explained as an investment in a research programme that promised very high payoffs if it were to succeed.

It is here, of course, that the difference between broad reductions—that is, part-whole explanations without remainder—and incomplete part-whole explanations is crucially important. Even partial explanations are good *explanations* insofar as they provide explanatory *insight*. For example, the kinds of microexplanations whose accumulation in the twentieth century Papineau alludes to are what Bechtel (2006) has called *mechanistic explanations,* which he points out are much weaker than broad reductions. Over the course of this work, Bechtel examines, in detail that Papineau does not, the kinds of explanatory insights provided in a number of such areas. But what they do not provide is any kind of knock-down metaphysical grounding for claims of metaphysical supervenience. While the scientist might well regard the incompleteness of existing microexplanations as an agenda for further research, the partial explanations in their own right are already important contributions to scientific knowledge. But for the metaphysician, without broad reductions in hand, claims of metaphysical supervenience still smack of something of an article of faith rather than something that has been satisfactorily established.

4.2. The Epistemic Status of Reductive Physicalism

Yet (as Pascal might admonish Descartes) faith need not be without its reasons, even if these are not demonstrative. Reductionism about the mind has been bolstered by two types of arguments, sometimes made explicit and sometimes merely gestured at. The first of these is an aprioristic and normative argument; the second is inductive.

4.2.1. The Normative Argument(s)

The normative argument can be rendered as follows:

1. It is a rational norm governing the sciences that the (true) claims of the special sciences must be such that they could in principle be derived by a kind of axiomatic reconstruction whose axiomatic base consists entirely of assertions cast at the level of basic physics.

2. Claims about mental phenomena are postulates of special sciences.
3. Therefore, they must (if true) in principle be derivable from truths of basic physics.

At least some eliminativists share with reductionists the normative claim (1), but deny (2):

1. It is a rational norm governing the sciences that the (true) claims of the special sciences must be such that they could in principle be derived by a kind of axiomatic reconstruction whose axiomatic base consists entirely of assertions cast at the level of basic physics.
4. Claims about mental phenomena cannot be derived from truths of basic physics.
5. Therefore, claims about mental phenomena are not *true* claims, and must involve erroneous postulates.

Indeed, if one accepts (1), one is faced with a kind of *forced choice between reductionism and eliminativism*: claims of the sciences of the mind must either be derivable (in principle) from claims cast at the level of basic physics, or else they must be regarded as false and their theoretical postulates eliminated.

The kind of aprioristic normative claim represented by (1) was typical of the Positivist and Logical Empiricist project, which sought less to describe the actual practice of the sciences than to regiment them into an "acceptable" logical form, which in turn served as a kind of normative litmus for the status of any given scientific claim or framework.[3] This philosophical attitude toward the sciences, however, is probably the element of the Positivist program that has been most widely rejected by contemporary philosophers of science. It is widely accepted that the special sciences are "autonomous," both in the sense that local explanatory successes are self-justifying and are not held hostage to reductive integration with more basic levels of explanation, and in the sense that they employ diverse methodologies and forms of explanation and postulate different types of entities. If we are to regard accepted scientific successes as the standard by which philosophical accounts of science are to be judged, rather than the other way around, we cannot accept something like (1) on *a priori* grounds, and indeed it seems to be a principle we would have to reject if we are to do justice to successful work in the special sciences.

And if there is no viable *general* normative principle like (1) to motivate the argument, it is hard to see why one should adopt such a principle in the *single* case of *mental* phenomena. To hold the sciences of the mind to such a standard would be to hold them to a much higher standard than we hold the other special sciences. While there may be a few metaphysicians who are inclined to draw eliminativist conclusions about the inventories of the special sciences generally (e.g., van Inwagen 1993), very few contemporary philosophers of science are tempted to see a forced choice between reducing or eliminating the entities of biology or other special sciences. And if there is no such forced choice generally, it is hard to see what could motivate it in the special case of psychology or other sciences of the mind (compare Stich and Laurence 1994; Baker 1995).

Of course, the mere (sociological) *fact* that philosophers of science presently embrace the autonomy of the special sciences does not itself entail that they are *right* to do so. It is conceivable that future generations will return to the kind of *a priori* normative approach practiced by the Positivists, and might do so on the basis of good reasons that respect the legitimate successes of the special sciences. But for this to be a serious objection, someone would have to actually present such reasons; and to the best of my knowledge, no serious proposal of this type is presently on the table among philosophers of science. In short, the normative argument for reductionism seems to be dead in the water.

4.2.2. The Inductive Argument

While the normative argument enjoys little support today, it is much more common to hear some version of an inductive argument for reductionism. Such an argument might be rendered as follows:

6. The collective evidence of modern science reveals that the phenomena of the special sciences are, in general, subject to unification through broadly reductive explanations.
7. If a principle (such as reducibility of the special sciences) applies broadly in the mature natural sciences, it is reasonable to suppose that it will prove applicable to the sciences of the mind as well.
8. Therefore, it is reasonable to suppose that mental phenomena will prove to be subject to broadly reductive explanation in terms of the physical sciences.

A similar argument can be marshaled for eliminativist conclusions:

6. The collective evidence of modern science reveals that the phenomena of the special sciences are, in general, subject to unification through broadly reductive explanations.
7. If a principle (such as reducibility of the special sciences) applies broadly in the mature natural sciences, it is reasonable to suppose that it will prove applicable to the sciences of the mind as well.
9. Therefore, we should reasonably expect of the special sciences that their phenomena will either prove broadly reducible or else end up as candidates for elimination.
10. Mental phenomena are not broadly reducible.
11. Therefore, they may reasonably be considered candidates for elimination.

The most glaring problem with the inductive arguments is that, in light of developments in post-reductionist philosophy of science presented in chapter 3, premise (6) seems to be false. There are, of course, very powerful part-whole explanations of phenomena in the special sciences, including psychology. But they are generally partial explanations rather than explanations without remainder, and hence do not rise to the level of broadly reductive explanations. In

particular, they do not carry the force of metaphysical necessity. If even the supposed reductions of thermodynamics, evolutionary biology, and classical genetics are not broad reductions, (6) would seem to be false. And even if one might show that there are a few examples of broad reductions to be found, this does not amount to the kind of general principle required by (6). If there are *only a few* known broadly reductive explanations after four hundred years of intense and fruitful scientific inquiry, and what we find in other cases is an abundance of partial explanations and explanatory gaps, we are not entitled to project any grand inductive consequences that support reductionism.

Premises (7) and (9) are also quite dubious. It is part and parcel of the general acceptance of the autonomy of the special sciences that one ought not to take features of the form or methodology of one scientific domain, or of the relations between any two of them, and expect that things will work in the same way with other domains or combinations of domains. As writers like Toulmin and Suppes noted (see quotes in chapter 3), the ability to come close to broadly reductive explanations in mechanics merely shows how unusual a science mechanics is, and does not license inductions to the special sciences. The inductive arguments for reductionism and eliminativism, like the normative arguments, turn out to rely on premises that seem quite implausible in the wake of post-reductionist philosophy of science.

4.3. ''New Wave'' Reductionism

Several contemporary writers have acknowledged the failure of broad reductionism, and yet have sought to maintain some notion of ''reduction'' that is only slightly weakened. John Bickle, in particular, has developed ideas of ''new wave'' (1998) or ''ruthless'' (2003) reductionism. Bickle's discussions, moreover, are developed in admirable dialogue with detailed case studies. Bickle's new-wave reductions are still very powerful explanations, but depart from broad reductions in a number of significant ways. I shall single out three of these. First, Bickle's metatheoretical framework allows for theories and models to be couched against a variety of types of implicit background assumptions. Second, Bickle's new-wave reductions involve *contingent identifications* of the entities picked out by descriptions in different theoretical vocabularies rather than metaphysically necessary type-identities or even necessary one-way type implications. Third, the reductions in question are what Bickle describes as ''token reductions,'' which is to say that they apply at the level of individual objects or phenomena rather than as wholesale reductions of a theoretical vocabulary or of the types picked out by that vocabulary.

It will be clear in chapter 7 that I think that the first move, of acknowledging the many ways scientific models are idealized and hence cannot be well represented by universally quantified claims in predicate logic, is both right-headed and crucially important. The third, the notion of ''token reductions,'' I am not quite sure what to make of. It is my impression that it is supposed to suggest something more than the second claim, of contingent identity, but

I am not certain how to understand a notion of "reduction" that applies to tokens rather than types. What is most important for our purposes, however, is the use of contingent identities. One might think that the use of the notion of "contingent identity" reflects having missed Kripke's lesson that identities are always necessary. However, I think this would be unfair. Kripke's work makes a case that the identities of *particulars* are necessary, and makes a more limited case for the necessity of identity for a very *limited range of kind-terms*, the so-called natural kinds. I think that there are reasons for suspicion of this latter claim, and, more important, that subsequent writers have sometimes invalidly extended it beyond the bounds of legitimate natural-kind terms. What I think Bickle means, however, is not that the identity of individuals is contingent, but rather that the fact that something is an A-instance (for the relevant classes of properties) does not make it metaphysically necessary that it be a B-instance, even if the identification of the two is useful for purposes of scientific theory. Indeed, it may be more than useful: it may point to an important empirical regularity. It may be that Bickle's "contingent identity" is doing the same work that notions of "natural necessity" are doing for other writers, such as Chalmers (1996), while avoiding the problems that may arise from treating empirical regularities as weak modal notions (difficulties that are, as I argue later in this chapter and in chapter 8, considerable).

The crucial difference between Bickle's "reductions" and broad reductions (which are more or less what he calls "classical reductions") is that they do not carry the force of metaphysical necessity. They certainly do not carry the force of conceptual adequacy, and hence do not *guarantee* metaphysical necessity in a fashion that is epistemically transparent. If there are metaphysical necessities to be had here, they are epistemically opaque to us. Bickle sidesteps this issue by treating them as contingent. This may very well be a far superior reflection of the metaphysical commitments of the sciences than the broadly reductive model. Indeed, I think there is definitely a need for *some* strong intertheoretic relation that falls short of metaphysical necessity here, whether it is contingent identity, nomological necessity, or some other notion. However, both contingent identity and nomological necessity debar one from proceeding from such weaker notions of "reduction" to the metaphysical conclusions that reductionists have traditionally drawn: namely, that the "reduced" system is metaphysically supervenient upon the "reducing" system. Thus, while Bickle is putting the *word* 'reduction' to a good use—indeed, a use that, as he suggests, better reflects how the word is used by scientists—it is important to recognize that, even if his analysis is correct, it does not license claims of metaphysical supervenience of mind upon brain. (Likewise, claims of nomic necessity posit nomic relations between kinds that are weaker than metaphysical supervenience.) Indeed, it is not clear that, at this level, Bickle and I really disagree, except in rhetorical strategy. He wants to reclaim the word 'reduction' for a good use; I wish to salt the earth on which it grew so nothing ever grows there again. His position and mine are, in some sense, notational variants, at least up to a point.

However, in his 1998 book, Bickle seems to wish to put the notion of contingent identity in the service of supporting token physicalism. It is true

that his account would render the intertheoretical linkages he describes *consistent* with token physicalism. But it is far less clear that there is reason to describe these relations as token identities as opposed to something else. Whenever there is an empirically robust linkage between properties that can be interpreted as (i.e., consistent with) a token identity, there are always alternative interpretations available. For example, they could be separate properties that are nomically related as cause and effect, as Descartes suggested for mind-body relations. Or they could be separate properties that are nomically related as effects of a common cause. They also admit of other interpretations, such as occasionalist or dual-aspect accounts.[4] In short, I find nothing in Bickle's analysis of the cases that should lead us to prefer his thesis of token physicalism, simply on the grounds of the science, to these alternatives, which have very different metaphysical commitments. (Nomic causal relations, at least, are quite respectable and widespread in the sciences.) So far as Bickle's case studies show, the science itself is quite neutral between competing metaphysical interpretations.

Bickle himself seems to have reached a similar conclusion in his 2003 book. There he disavows traditional metaphysical projects in philosophy of mind entirely in favor of an empirically driven philosophy of neuroscience. At one level, this may be seen as simply an entirely reasonable research strategy: to pursue a "hands-on" philosophy of neuroscience in its own right, apart from issues in metaphysics, much as philosophers of physics or biology tend to pursue their studies without reference to metaphysical issues. Indeed, in chapter 9 I offer some deep reasons to be suspicious of the project of necessitarian metaphysics. However, it is also, importantly, an admission that, however successful Bickle's analysis of the relations one actually finds between the sciences, it does not yield results that can be put to use in the service of a particular metaphysical view, such as physicalism. "Ruthless reduction" is thus not a revival of the broad reductionist project, but a repudiation of it. At the level of explanation, it gives up on the project of making type-relations completely epistemically transparent. And it disengages explanation from the metaphysical project entirely.

4.4. Nomological Necessity

A second strategic retreat from broad reduction is found in the notion of "nomological necessity" or "natural necessity." These notions are generally contrasted with metaphysical necessity. If B→A is metaphysically necessary, then all B's are A's in every possible world, and it is ultimately contradictory to assert B and deny A. Because laws of nature do not seem to be truths of reason, as Descartes and other Rationalists briefly supposed, their denial is not self-contradictory, and there are presumably possible worlds that have different laws, or perhaps even no laws at all. Yet at least some laws, such as those of fundamental physics, are supposed to have universal scope within the actual world, and additionally to be counterfactual-supporting. One way of

interpreting this is to view law claims as a type of modal claim with a force more restricted than that of metaphysical necessity, and yet greater than that of material implication, and to apply to the actual world plus all of those other possible worlds to which the same laws apply. Both physicalists and dualists have attempted to make use of notions of nomological or natural necessity, though to my mind the retreat to merely nomological relations is itself a decisive step away from physicalism, if not necessarily from materialism. And so this topic might be better taken up in the chapter on dualism. However, because some physicalists seem to view it as a way of salvaging their position in nonreductive terms, I shall explore it here.

I think there are deep problems in working out a viable account of laws in modal terms. Some of these will need to await development in chapter 9, as they depend on my own Cognitive Pluralist views. But others can be briefly developed here. One of these is signaled by Bickle's right-headed observation that scientific theories, laws, and models are generally hedged by a variety of background assumptions and idealizations. While Bickle's examples are taken largely from the special sciences, the point has been made forcefully by philosophers of science like Nancy Cartwright with respect to physical laws as well. A given law, such as the gravitation law, may in some sense "have universal scope" in that it is always "in play"; but what it does *not* do is tell us how things always actually *behave* in nature. Objects do *not* actually behave as the gravitation law would predict when there are other causal factors, such as aerodynamics or electromagnetism, at work as well. A paper airplane does not fall like an identical piece of paper of equal mass crumpled into a tight ball. A metallic object does not fall if it is in the field of a suitably strong magnet. Even laws like the gravitation law, if taken as universal claims about how objects actually *behave*, have exceptions. Indeed, because multiple forces are *always* "in play," it is likely that even many of the most basic laws have *nothing but* exceptions, if interpreted as universal claims about actual behavior.[5]

Some laws *can* be taken as universal if they are supposed to express something else, however, such as causal capacities (Cartwright 1989, 1999) or forces (Horst 2004). To the extent that we can factor and sum these component forces through vector algebra, they provide the kind of fully determinate kinematic descriptions needed to ground the counterfactual force of laws. However, Cartwright argues that, for many of the types of laws we have, we do not have adequate ways of factoring and summing forces in all instances, in part because of the implicit background assumptions and idealizations already noted. If this is so, then it is not clear that the kinds of laws the sciences actually yield are adequate to the task of yielding the kinds of determinate values for counterfactuals required by a modal analysis of laws. That is, it is not clear that the kinds of regularities we find expressed in real scientific laws, theories, and models are sufficient to ground the kind of "nomic supervenience" desired on logical and metaphysical grounds, or required for a modalized interpretation of laws. Aspects of this problem are discussed more fully in chapter 9.

A second problem about the notion of nomic necessity is that it is unclear how to develop it in a noncircular way. There would seem to be two strategies

for cashing out laws in terms of modal logic. One is to treat laws as constructions out of the empirical generalizations true at world W, both of events that occur at that world and the truth-values of counterfactuals at that world. That is, take the basic facts as including those true at worlds plus the counterfactuals true at worlds, and then treat laws as expressions of the regularities found in these. This would be a neo-Humean strategy modeled on the interpretation of laws in terms of empirical generalizations that can be cast in terms of quantified predicate calculus, only extended to embrace counterfactuals as well. The second strategy is to identify the laws independently of our characterization of possible worlds and assignments of truth-values to counterfactuals, and treat the latter as derivative from these laws. This would be to treat laws as basic features or properties at worlds that are independent of or prior to the actual and counterfactual kinematic features (e.g., to treat laws as expressing real occurrent forces).

The neo-Humean strategy has the advantage of having closer parallels to the development of standard modal logics, in that the modalized claims (e.g., that a proposition is metaphysically or nomically necessary) are model-theoretic truth-functional constructions out of local assignments of truth-values to propositions at worlds. This strategy, however, faces several obstacles. First, unlike claims of metaphysical necessity and possibility, claims of nomic necessity and possibility at a world W cannot be read off the values of noncounterfactual propositions at W because the nomically impossible need not be self-contradictory. One must either (a) treat each counterfactual value as a brute fact, and the global regularities as emergent from these, or (b) treat some kind of ''nearness'' metric between worlds as brute. Neither of these approaches seems to give laws the kind of grounding they require, and indeed both seem almost arbitrary. Either might suffice for a *logical reconstruction* or *formal regimentation* of law talk suitable for preserving truth-values in inferences, but neither distinguishes accidental generalizations from the deep causal invariants that laws might reasonably be supposed to express, nor explains the division of sets of worlds in a principled way. There is nothing comparable to the principle of noncontradiction to provide sufficient reason for each occurrent fact or counterfactual truth underwritten by a law at W. It seems sensible to suggest that there might be possible worlds, for example, that are exactly like the actual world, including the physical counterfactuals true at the actual world, but whose regularities are the result of occasionalist divine causation. Such worlds would, of course, be different with respect to the truth-values of *some* propositions, such as those involving God, but they would be *physically* indistinguishable, so long as laws express only actual regularities and not underlying *causes* such as forces. This, however, pushes us in the non-Humean direction of classifying nomic claims on the basis of *how events are caused* rather than simply in terms of the values of actual and counterfactual claims about the kinematics of physical objects and events (or, alternatively, of treating propositions about nonphysical entities and events as necessarily involved in specifying the physical facts and laws at a world).

However, if we assume that laws pick out deep invariants that are independent of or prior to facts about objects and events—for example, if we

interpret the gravitational law as expressing a fundamental gravitational *force*—we are faced with problems as well. If laws are understood as expressing forces, two worlds might have all of the same *physical* forces, but differ in that one has nonphysical forces or free-agent causation as well. At least some such worlds would differ with respect to what events occur at them, and they would certainly differ in the values of the subjunctive conditionals true at them. Likewise, if consciousness and intentionality are not metaphysically supervenient upon physical phenomena, one world might be a zombie world and another have beings with consciousness and intentionality, even if these were purely epiphenomenal. As a consequence, nailing down the physical facts and forces operative in a world W does not nail down *all* the contingent facts true at W, nor the values of counterfactuals at W. Nomic necessity, thus construed, cannot be put to use in adjudicating metaphysical disputes about the relation of mind and body.

There are at least two ways that the notion of nomic necessity might be augmented to handle such problems. The first is to add *negative* clauses to the description of worlds (Chalmers 1996; Polger 2003). One might, for example, treat propositions to the effect that there are (or are not) nonmaterial substances, causally efficacious nonmaterial properties, or instances of free-agent causation at a world W as partially constitutive of what is nomically necessary/possible at W, and as needed to determine whether two worlds W and W^* have the same laws. Again, however, this debars the resulting notion of nomic necessity from doing any work in metaphysical disputes. Materialists and dualists may very well agree on the list of physical forces at work in the actual world, but disagree on whether there is also consciousness, intentionality, or free-agent causation. If the nomically possible worlds are those that have all of the same *physical* laws, these may yet differ in terms of other properties, and in the case of free will, even with respect to the truth values of propositions about events and counterfactuals. If it is not metaphysically necessary that mental properties supervene upon physical properties, and laws express forces or other regular causal contributors, worlds may be nomically equivalent and identical in initial states, and yet diverge in their actual histories. Additionally, nailing down the physical laws does not suffice to determine which sort of world we live in. But if nomically possible worlds are more restricted and consist only in those that share, not only the same physical forces, but also the same "psycho-physical" laws linking matter to consciousness and intentionality and stipulations about whether there is free-agent causation at a given world, then what is nomically possible (at the actual world) depends on whether there is nonmetaphysically supervenient consciousness, intentionality, and free will in the actual world. As these are precisely the questions that are in dispute within metaphysics of mind, such a notion of nomological necessity can do little to help us to adjudicate them (though they might help to render more perspicuous the possible worlds at which physicalism is true and those at which it is not).

A second strategy, akin to the last move, is to treat psycho-physical relations (e.g., the relation between pains and C-fiber firings in a given species) as

an *additional type of law* that needs to be specified in order to pin down what is nomically necessary/possible (Chalmers 1996). This, of course, still leaves open the problems with free will, as free-agent causation is by definition not determined in even a lawlike way by prior events. But even with respect to Chalmers's suggestion that there might be brute "psycho-physical" laws relating brain states to consciousness, there are potential problems. First, it is not clear what kinds of "laws" these are supposed to be. They are *not* like the type of "psychophysical" laws found in scientific psychophysics, such as the Weber-Fechner laws, which report merely empirical generalizations about relations between stimuli and percepts. (I use the hyphenated spelling "psycho-physical" to distinguish the two usages.) Nor, if Chalmers is right, do they report causal mechanisms. Rather, they are contingent, but brute, relational facts obtaining between particular kinds of objects or phenomena, both in the actual world and also in at least nearby possible worlds. They are thus, at one level, on a par with the postulation of brute fundamental forces like gravitation, an interpretation that Chalmers seems to endorse.

This strategy, unlike the previous ones, does not seem to me to be metaphysically neutral: it treats mental and physical properties as quite distinct and requiring a brute law to link them. It thus seems to require at least a property dualism, though perhaps not a substance dualism. However, when laws are interpreted in terms of forces, they are generally understood to express actual and potential causal relations, and it is not clear that Chalmers's "psycho-physical" laws should be understood this way. To be sure, both epiphenomenalist and interactionist accounts are available that *do* treat psycho-physical relations as causal in character: the stimulation of my retina sets in motion a chain of events that causes me to have an experience with a particular phenomenological character, or (on the interactionist though not the epiphenomenalist interpretation) my desire to perform an action causes my body to move so that it is thereby performed. But it is not clear that all psycho-physical regularities *require* a causal construal, and indeed it seems more natural to think of at least some of them—for example, the relation between the firing of C-fibers and the experience of a painful sensation—as *not* reporting a causal interaction between two separate events, even if they report relations between two separate property instances. It is not clear that we should say, for example, that a C-fiber fired at *t, and then as a causal consequence* I felt a pain at $t + \partial$. Indeed, it is not clear *what* we should say about such a relation. The relation seems to be quite *sui generis,* and not to fit into the molds of identity, constitution, or causation.

Moreover, however we interpret them, psycho-physical laws would seem to run the risk of having the result that the "forces" postulated are not independent of other fundamental forces. There are no gravitational-electromagnetic laws, for example. Gravitation and electromagnetism both influence how objects behave, but the two types of *forces* are independent. And if psycho-physical laws are supposed to be fundamental in the sense that gravitational and electromagnetic laws are supposed to be fundamental, this would seem to require independence. But for there to be psycho-physical laws, this would seem to violate the requirement of independence.

Of course, even if Chalmers's approach works, it should yield no comfort to the erstwhile reductive physicalist. If facts about the mind are underdetermined by the physical facts, then physicalism is false. Brute psycho-physical laws, interpreted as deep facts about the world rather than mere empirical generalizations, seem to require a property dualism that has real bite to it.

4.5. Reduction and Metaphysics

While reductive explanation has been posed as a problem for philosophy of psychology (i.e., as constituting a need to "vindicate" the sciences of the mind by showing them to be reducible), it has played a more prominent role in the service of arguments about the metaphysics of mind. Reductionists have used the assumption that mental phenomena are (or must be) broadly reducible in the service of arguments for reductive physicalism, and eliminativists have argued that the irreducibility of mental phenomena undercuts, not only their status in particular special sciences, but also their ontological credentials.

Scientific pluralism, however, threatens this connection between reducibility and ontological status in a very fundamental way. First, if reducibility is a criterion for ontological legitimacy, it is not only intentionality and phenomenology that are threatened, but phenomena of other sciences as well, such as organisms, species, metabolic processes, temperature, and molecules. Indeed, it would seem that we are presented with a forced choice. On the one hand, one might bite the bullet with van Inwagen and hold that, in some privileged sense of "reality" and "existence," only the most basic particles can really be said to really exist. On the other hand, one might retain one's commitments to the reality of other things at the price of rejecting a connection between reducibility and ontological legitimacy. There are, of course, familiar ways of doing this, such as Pragmatism and Kantian idealism. But *however* one approaches the matter, it looks as though reductionists and eliminativists must either sacrifice deeply entrenched ontological commitments or else admit that a broad variety of things can be in good ontological standing even if they are irreducible.

Second, scientific pluralism raises deep questions about just what materialism and physicalism amount to. It is easy to miss the difficulty if one is content to rest with abstract formulations, such as that materialism is the doctrine that everything that exists is a material object, or physicalism the doctrine that everything that exists is nothing but a simple or complex object of the sorts that appear in the ontology of physics. But what is meant by this "nothing but"? The original, broadly reductionist understanding is both familiar and intelligible: it means that every (legitimate) type of phenomenon, event, object, or property can be reconceived and completely accounted for in physical terms. This formulation seems clear enough, but if broad reductions are rare or nonexistent, it is also false. If 'materialism' and 'physicalism' are to continue to be meaningful labels that describe live philosophical options, we must provide characterizations that are compatible with both the reality and the irreducibility of the objects of the special sciences.

One option might be to claim that every (real, legitimate) object is, among other things, a physical object, even if its physical-level properties do not necessarily provide a base for the metaphysical supervenience of its other properties upon them. (This seems to be more or less the course explored by P. F. Strawson [1959].) This, of course, has the usual problems in accounting for abstract objects like numbers, but as that problem is faced by *any* form of materialism or physicalism, it will not receive special attention here. More problematic is the fact that it would admit such views as Aristotelianism and Spinozism into the materialistic fold. All of Aristotle's substances are hylomorphic unities, and even Spinoza's God is rightly describable in material terms. Indeed, proponents of many forms of Christianity believe in an everlasting life in a resurrection body rather than an ethereal existence as a Platonic nonmaterial soul. And indeed many of the Scholastics held that even angels are individuated through a special type of matter. Such a view would exclude Cartesian souls and a God who preexisted all matter (though it is not clear on what principles one would base such an exclusion), but would admit a surprising array of things as "material."

4.6. Conclusion

Reductionism and the forms of eliminativism that are essentially the flip side of reductionism are severely threatened by theory pluralism in philosophy of science. Both normative and inductive arguments for these positions depend crucially on the assumption that intertheoretic reductions are commonplace at the junctures between the natural sciences, but this assumption is undercut by post-reductionist philosophy of science. Attempts to save reductionism by weakening the notion of "reduction" in play result in accounts that no longer view the mental as metaphysically supervenient upon the physical, thus violating the assumption of physicalism. Short of a wholesale revolution in philosophy of science, and perhaps in the sciences themselves, that revitalizes the broad reductionist program, scientific pluralism would seem to deal a mortal blow to both reductionism and eliminativism in philosophy of mind.

This page intentionally left blank

5

The Explanatory Gap and Dualism Reconsidered

If scientific pluralism is bad news for reductionism and eliminativism, one might expect it to be correspondingly *good* news for dualists and other traditional friends of the explanatory gap. This, however, turns out to be only half true. On the one hand, scientific pluralism does indeed go a long way to silence concerns about the reality of explanatory gaps. But on the other hand, it also puts pressure on the dualist to count beyond two and to embrace a more radical pluralism.

5.1. The Status of Explanatory Gaps

Recent discussions of the explanatory gaps between mind and body have generally involved three assumptions:

1. We presently find such gaps.
2. At least some such gaps are principled and abiding.
3. The psychological gaps are unique: one does not find similar gaps in the natural sciences.

It is the tension between (1) and (2), on the one hand, and (3) on the other, that makes the psychological gaps such a fascinating philosophical (and scientific) problem. To the extent that one embraces the claim that the psychological gaps are unique (3), one might reasonably feel a sort of suspicion about the assumption that they are principled and abiding (2). If, in general, de facto explanatory gaps (found at any time during the progress of scientific understanding) tend to be closed eventually, we have reason to think that the psychological gaps are not principled and abiding. And of course the sciences of cognition *are*

progressing toward shortening the breadth of some explanatory gaps, and so one might suppose that things will eventually work out.

On the other hand, to the extent that one takes both (2) and (3) seriously, one is faced with a striking problem: Why is it that the mind is separated from the world of nature in this way? The situation is one that has variously motivated dualism, eliminativism, and the search for viable nonreductive forms of materialism. It is unclear that there would be much motivation to be anything but a reductive materialist without (2). Without (1) and (3) . . . well, it is hard to know just how problems in philosophy of mind would shape up.

Scientific pluralism provides a powerful vindication of the psychological gaps. It does so in an indirect way: it does not provide any direct new evidence about *those* gaps, but rather provides reason to believe that there is *nothing unusual or problematic about explanatory gaps in general*, even principled and abiding ones. If we seemed to find such gaps in only one place, and indeed only with respect to the objects of some of the least developed and most complicated sciences, we would have at least some reason to suspect that the psychological gaps are not real and abiding. But if the gaps we find there are indeed the sort of situation we routinely find at the boundaries between scientific theories, we have no principled reason to treat the status of the psychological gaps with suspicion. Indeed, at some level, *they are precisely what we ought to expect*. To find that consciousness is *better* explained by neuroscience than evolutionary biology by molecular genetics would, after all, be quite a surprise. This does not mean, of course, that we should stop trying to find whatever intertheoretic explanations of mental phenomena we can find, any more than we should stop seeking physical explanations of chemical phenomena or explanations of phenotypic features in molecular biology. What it means is merely that the mere existence of explanatory gaps in one special science does not by itself constitute a difference from what we find throughout the rest of the universe as understood by the sciences, nor necessarily a unique and sexy philosophical or scientific problem.

But of course, if this is so—if explanatory gaps are commonplace—then it is not clear why the *psychological* gaps ought to be such a big deal. There is obviously an important question of why there are, in general, such gaps in our understanding. But thus far we have no reason to suppose that the reasons for the mind-brain gap are fundamentally different from those for, say, the evolution–molecular genetics gap. And, perhaps most troubling for the dualist, almost no one is inclined to use the other explanatory gaps as a basis for arguments in ontology. No one, for example, has suggested that we are faced with a "gene/species dualism." Yet the dualist wants to make precisely such arguments with respect to the mind.

5.2. The Dualist's Dilemma

Dualists have, in modern times, argued for their views primarily on the basis of what we now call explanatory gaps. In a seminal discussion, Descartes offers

such an argument in Book V of the *Discourse on the Method*, where he claims that, whereas *all* processes found in nonhuman animals can be replicated (and hence explained) mechanically, there are at least two faculties humans possess—reasoning and language—that cannot be mechanically replicated, and hence require the postulation of something nonmechanical (and hence, given Descartes's mechanistic conception of physics, nonphysical) to explain them. This is not, of course, Descartes's only argument for dualism, but it is in some ways the most successful and fertile one: *successful* because versions of it are still to be found in the literature, and *fertile* because it in many ways defined an agenda for would-be physicalists, and might, for example, be seen as the basic agenda driving projects like artificial intelligence today.

Contemporary proponents of dualism, notably David Chalmers (1996), have argued in a similar fashion, from the existence of principled and abiding explanatory gaps to a form of metaphysical dualism. Chalmers's dualism is a dualism of properties rather than of substances, but the basic argumentative strategy is the same. I concentrate here on property dualism, but the same concerns will apply, *mutatis mutandis*, to substance dualism as well. The Main Argument for dualism can be put as follows:

D1. There are mental properties M for which there is no set of nonmental properties N such that it is possible to give a CAE of M in terms of N (*the Explanatory Gap*).

D2. If it is impossible to give a conceptually adequate explanation of A in terms of B, then it is not metaphysically necessary that B implies A (*Negative EMC*).

D3. There are mental properties M for which there is no set of nonmental properties N such that it is metaphysically necessary that N implies M (*Failure of Metaphysical Supervenience*).

It is worth noting that D3 actually does not end with a conclusion that is exactly an assertion of dualism. D3 is a denial of physicalism, in that it claims that there are mental properties that are not metaphysically supervenient upon any nonmental properties. However, it leaves open the possibility that there are also *other* types of properties that are both nonmental and nonmaterial. The dualist is generally concerned primarily to argue that the number of fundamental property types (or the number of substance kinds) is *greater than 1*, not that it is *fewer than 3*. The reason for this is that many prominent dualists, including both Descartes and Chalmers, have explicitly endorsed broad reductionism outside of the realm of the mind: they have supposed that everything else in the world is broadly reducible to, and hence metaphysically supervenient upon, basic physics. Dualists have generally *assumed*, along with materialists, that there is a single type of substance or property that may be designated "material," and then argued that one must recognize exactly one additional type of substance or property, amounting to a grand total of 2.

It is here that scientific theory-pluralism presents a problem for the dualist. If we accept both Negative EMC *and* scientific pluralism, we can run

variations of the Main Argument for *every irreducible type of object, property, or system*. I am not sure just how many types of irreducible objects and properties there are, or even whether there is a single and determinate answer to this question, but we would surely be required to count to a number greater than 2.

This would seem to present the dualist with a nasty dilemma. On the one hand, he can hold on to Negative EMC and accept the result that the ontological position implied is not dualism but a pluralism of a much higher ordinality. I suspect that many dualists would find this result preferable to materialism, but it is nonetheless a significant move away from traditional dualism. And as we shall see in chapter 7, such a radical ontological pluralism has some very counterintuitive consequences of its own.

On the other hand, the dualist might hold on to dualism by abandoning Negative EMC as a general principle, thus barring the move from scientific pluralism to a more radical ontological pluralism. Such a dualism would be *consistent*. However, it would be much weakened in its evidential status. Because Negative EMC is a crucial part of the principal argument in the dualist's arsenal, to abandon it is to surrender the main argumentative reason for embracing dualism in the first place. It might be possible to revive other arguments, such as Descartes's Real Distinction argument from the *Meditations*, or arguments from the (supposed) simplicity of the soul stemming from Plato's *Phaedo*, or experiential arguments based on near-death experiences or extrasensory perception, but these are generally viewed as being in much weaker standing on the current philosophical scene.

This seems to be a destructive dilemma. On the one hand, the erstwhile dualist can retain Negative EMC at the expense of abandoning dualism. On the other hand, he can abandon Negative EMC at the cost of leaving dualism essentially without the support that has gained it some philosophical currency. Such a dualism would still be tenable, in the sense of being internally consistent and consonant with the data, but it would be as much a standpoint of faith as post-reductionist materialism.

5.3. Why is This Gap Different from All the Other Gaps?

But whereas post-reductionist philosophy of science might provide a decisive argument against reductive physicalism and eliminativism, the dualist has at least one, more promising option to explore that would allow him to hold on to both dualism and some version of the Negative EMC. The strategy begins with an intuition: Many people who acknowledge an abundance of explanatory gaps in the form of failures of broad reducibility nonetheless experience the intuition that *the psychological gaps are somehow different from, and deeper than, the others*. Indeed, I myself experience such an intuition quite strongly. Like any intuition, this one might turn out to be illusory. But it might also turn out to point to something real, and pursuing it might lead to a substantial deepening of our understanding both of mind-body relations and of explanatory gaps in general.

I take this to represent an important research agenda for would-be dualists. Whereas broad reductionism cannot be salvaged without a wholesale change in our natural sciences, one in which the gaps we find everywhere are closed by broadly reductive explanations, dualism might be saved merely by showing that there is some special feature of the psychological gaps that is not found elsewhere, and that Negative EMC can be rightly applied only to those that have this special feature. As I have not personally been substantially more attracted to dualism than to materialism for quite some time (roughly since the time I experienced a serious knee injury and found myself experiencing the clear intuition that it was *I*, and not merely my body, that was injured), I am not motivated to find such a solution myself, and indeed in chapter 8 shall offer an alternative explanation of the intuition that the psychological gaps are special. However, it is worth pursuing one line of thought that might be attractive to dualists, even though I think that ultimately it is not adequate to the task at hand.

5.3.1. The Argument from Lack of Suitable Candidate Explainers

In *Symbols, Computation and Intentionality* (Horst 1996), I argued that intentionality and phenomenology are not reducible to computation or other physical processes on the grounds that our theories of computation and of physical systems do not provide even *candidate explainers* for such things as meaning or subjectivity. To this list one might also add normativity, whether moral (Moore 1903) or semantic (Brandom 1994). Chalmers (1996) pursues a very similar strategy, on the grounds that reductive explanations are suited to explaining structural and functional properties, and *only* those sorts of properties, while the crucial properties of conscious phenomenology are neither structural nor functional in nature. Chalmers's arguments, and mine, were designed to move beyond the claim that we cannot, *at present*, reduce mental phenomena to physical phenomena to the claim that we cannot do so even *in principle*, because the supposed reduction base does not have the right sorts of resources. (One might take as a guiding analogy the example of first-order logic lacking the resources to construct modal logics out of it.) But such arguments might also provide a principled way of distinguishing the mind from the phenomena of other special sciences, and thus differentiating the psychological gaps from all the other gaps.

What Chalmers explicitly said, and I at least implicitly accepted at the time, is that all of the phenomena in the natural sciences *do* have at least candidate explainers—that is, that basic physics provides the kinds of resources that one could, in principle, use to provide broadly reductive explanations of phenomena in chemistry and biology, even if we do not always have such explanations at hand at the moment. It was not until 1997 that I began to be aware of the literature in philosophy of science that seemed to suggest that this assumption was in error: that we in fact have almost no broad reductions even in the natural sciences, and are not likely to get them in the future. On the pluralist picture, even ideally completed physical sciences would be gappy, and hence there are not really any legitimate candidate explainers available. The problem

is not merely that we have not yet found the candidates that will fit the bill, but that *none* of the candidates, known or unknown, would turn out to do so. But if this is the case—if there are many principled and abiding explanatory gaps within natural science—then it would seem to follow that physics does not really provide candidate explainers for many of the phenomena studied in the natural sciences either, at least if "candidate explainers" means "candidate reduction bases." And if this is the case, then the lack of candidate explainers for things like consciousness and intentionality does not itself set them apart from phenomena of, say, biology.

There is, however, something unsatisfying about this. In *some* sense, it must be true that, if A cannot be broadly reduced to B, then B must lack the resources needed to provide even candidate explainers for some features of A. Yet while we can in some sense just *see* that facts about fundamental physical objects are not the right kinds of things to explain phenomenological feels and normative principles, much as we can see that first-order logic does not have the resources to construct modal logic as a conservative extension, we nevertheless feel that they *should* be the sorts of things that can completely explain, say, chemical or biological phenomena. At the very least, the mismatch between the things to be explained and the potential explainers is not so intuitively evident in the physical sciences as it is in the case of phenomenology, intentionality, or normativity. The psychological gaps seem intuitively obvious, at least to some of us. The idea that there are abiding explanatory gaps in the natural sciences, by contrast, comes as somewhat of a surprise and calls for a fundamental reexamination of scientific metatheory. We are thus at least in a very different *epistemic* position with regard to the questions of whether physics provides candidate explainers for biological phenomena and whether it (or neuroscience) does so for mental phenomena.

My own assessment of this is that, with respect to the natural sciences, we suffer from something like a Kantian dialectical illusion that leads us to assume that they can be reductively unified. This interpretation is developed in part III. But it is hard for me to see how the dualist can accept scientific pluralism in the form of principled and abiding irreducibility of the special sciences and still argue that physics provides candidate explainers for things that stand on the opposite sides of explanatory gaps. To do so, she must hold that the nonmental explanatory gaps, but not the psychological gaps, are really just artifacts of our current ignorance; that is, the dualist must hold on to hopes for the vindication of reductionist philosophy of science. Such a hope strikes me as quite unreasonable.

5.3.2. Partial Explanation versus No Explanation

The strategy that Chalmers and I took in the 1990s might be modified into a more promising form. One might think that all that the other gaps amount to is the fact that there are *some* features of the special sciences that cannot be arrived at through axiomatic reconstruction from basic physics. Of course, there may also be many features that *can* be derived in this way, and more

that can be explained in weaker ways, such as mechanistic explanations. And so the "gappiness" in question is merely a matter of *incompleteness* of explanation. It is not explanation *without remainder*. But these are comparatively small gaps, like cracks in the pavement. With consciousness, intentionality, and normativity, by contrast, the chasm seems much wider. For these phenomena, one might think that the physical sciences provide *absolutely no explanation*. Laplace's demon, given a complete description of the physical universe but no independent information about consciousness, intentionality, or normativity, would have no clue as to the existence of such phenomena. And *this* sort of gap is more like the Grand Canyon in its span.

Of course, I do not expect that every reader will share the intuition that there is a deep difference here, any more than everyone agreed with Chalmers and me when we raised arguments of this kind against reductionism in the 1990s. For my own part, I still feel the force of such intuitions very strongly and consider them important data to be explained, or else explained away. On the other hand, it is not clear that intuitions about the "breadth" of various explanatory gaps ought to lead to one particular metaphysical conclusion rather than another. We cannot, for example, simply modify Negative EMC to apply only to "*really* broad" explanatory gaps without making the notion of "breadth" much more explicit. I think that it is incumbent upon would-be dualists to try their hands at this project, and I supply my own cognitivist explanation of the difference between the various gaps in chapter 8. Here, however, I conclude by exploring one obvious strategy that I think is ultimately inadequate.

One way of interpreting the contrast between the psychological gaps and those in the natural sciences is to say that, in the former case but not the latter, *nothing at all* can be explained by the candidate reducing system. Classical particle mechanics may not be able to explain the temporal asymmetry of entropy, but it does explain a great deal about thermodynamics. Particle properties may not be able to explain all global properties because of quantum entanglement, and indeed entanglement seems to force us to say that even particle-level properties are not entirely local, but they still explain a great deal. But one might claim that neuroscience and other sciences of the mind explain nothing at all about phenomenology, intentionality, or normativity. And so we might restrict Negative EMC to the latter sort of case, where nothing at all is explained, thus blocking the argument from scientific pluralism to radical ontological pluralism.

I think there is an interpretation of this claim on which it is true and an interpretation on which it is false. What is true is that Laplace's demon, supplied with a full physical description of the world and its natural laws (excluding any psycho-physical laws), and nothing more, would have no basis on which to suppose the existence of phenomenological feels, intentional states, or norms. For all it knows, the world is populated entirely by zombies. The physical sciences provide no explanation, for example, of why seeing red has the precise phenomenological feel it has, or indeed why it has any phenomenological feel at all (Chalmers 1996; Jacobson 1997)—similarly, *mutatis mutandis*, for the intentionality and normativity of mental states.

However, it does *not* follow from this that neuroscience can supply *no* explanations of any sorts of phenomenological facts whatsoever. *Given* that our visual phenomenology is either realized by or caused by a particular type of neural structure, a great deal about the abstract shape of our phenomenological space follows quite straightforwardly: for example, that phenomenological color space of trichromats takes the form of the Munsell color solid, or that we can experience a phenomenologically pure yellow but not a phenomenologically pure orange (Horst 2005). Such explanations, of course, depend crucially on taking the existence of phenomenological states, and their realization in, or causation by, particular brain states, as *given*—that is, as a background assumption—and hence do not explain them. These explanations are thus "mixed" explanations, in the sense that they are based on assumptions both about phenomenology and about the realizing system or their neural causes. But mixed explanations are rampant in the sciences (e.g., the molecular biologist or developmentalist might take evolutionary history as a given). And they *do* provide partial explanations, and even very powerful ones. Indeed, in cases like human color vision, they provide explanations that are about as close to CAEs as one finds anywhere, as many psychophysical data (interpreted as facts about discriminative abilities rather than phenomenology) can be derived mathematically from known properties of the cone and ganglion systems in the retina (see Horst 2005 for an exploration of this topic). So it would be untrue to say that disciplines like neuroscience can explain *nothing about* phenomenology, even though they say nothing about why phenomenological properties are present in the first place. (One might suppose similar things can be said, *mutatis mutandis*, with respect to intentionality and normativity.)

How different is this from the situation in the natural sciences? One might yet make a case for a crucial difference by recalling the contrast between different types of CAE discussed in chapter 2. With pure CAEs, all of the terminology of the theory of the system that is being explained must be present in, or constructible from, that of the explaining system. Phenomenological vocabulary is not present in the vocabulary of basic physics, and (arguably) even Laplace's demon cannot construct it from that vocabulary. Likewise, it cannot derive "ought" from "is," and (I would claim) cannot construct intentionality out of physical or neural facts either. But by the same token, 'adaptation' and 'selection' are not in the vocabulary of basic physics, and if writers like Kitcher (1984) and Lewontin (1983) are correct, cannot be constructed out of it either. However, if 'adaptation' and 'selection' are functional notions, we might supply the demon with independent definitions of these terms, and then it might very well be able to look at the history of the universe, described in purely physical terms, and identify what phenomena should count as adaptations and instances of natural selection. It could, that is, provide a CAE in the form of *criterion filling* of why each particular counts as an adaptation or an instance of selection. By contrast, even if supplied with phenomenological notions, it could not tell whether organisms were conscious subjects or zombies, as the phenomenological kinds are not functional kinds. Here, one might think, is a crucial asymmetry between the explanatory gaps found in psychology and those found elsewhere.

While this distinction seems intuitively plausible, this is just the sort of intuition that the past several decades of philosophy of science have given us reason to be cautious about. We cannot just assume from the armchair that physics *does* provide the basis for criterion-filling CAEs of things like selection. We can justify this only by way of repeated and careful case studies of intertheoretic explanation in various sciences. This is a discussion that needs to take place, but it will require philosophers of mind to enter into closer engagement with "hands-on" philosophy of science. And dualism, in particular, cannot be argued (at least as an alternative to pluralism) apart from more general issues in philosophy of science.

5.4. Dualism without Necessitarian Metaphysics: The Best-Interpretation Strategy

Not all defenses of dualism have been based on arguments involving the intricacies of necessitarian metaphysics. William Robinson's (2004) *Understanding Phenomenal Consciousness*, for example, adopts the strategy, not of arguing that the psychological gaps *entail* dualism, but that dualism is the *best interpretation* of such gaps when compared with such traditional competitors as reductionism, nonreductive physicalism, and eliminativism. Robinson's book is the most comprehensive defense of dualism I have encountered, and I have no real qualms with the claim that dualism compares favorably to these traditional alternatives.

But Robinson takes the view that the psychological gaps are unique and does not consider the problems arising from scientific pluralism. Nor does he compare dualism with the more radically pluralist views that are discussed in chapter 7. Nor has any other dualist, to my knowledge, undertaken the task of showing why dualism is to be preferred to pluralisms of a higher ordinality. Indeed, one might well think that at least some arguments to the effect that the psychological gaps are best interpreted by a dualist metaphysics could be adapted to produce arguments that the nonpsychological gaps are best interpreted by a more radical form of pluralism. Robinson's book addresses the comparative strengths of dualism and various forms of *materialism*. But a similar comparison of dualism and forms of *pluralism* has yet to be offered; without it, we have no reason to think that dualism is the *best* interpretation of *all* the evidence, but merely that it fares better than *materialism* in accounting for the specifically *psychological* gaps. And the argumentative strategy Robinson employs is one that pluralists might be able to turn to their advantage.

5.5. Conclusion

Whereas scientific pluralism deals what seems to be a mortal blow to reductionism and at least some forms of eliminativism, its threat to dualism is of a more limited and provisional nature. I have made a *prima facie* case that

scientific pluralism presents the dualist with an unpleasant dilemma: either forsake dualism for a more radical form of pluralism, or else abandon the Negative EMC, and with it the principal argumentative reason for embracing dualism in the first place. There is, however, a strategy dualists ought to explore that might provide a way out of this dilemma: namely, to show that the psychological gaps are different from the other gaps in a fashion that implies that Negative EMC is applicable in the case of phenomenology, intentionality, and normativity, but not in the cases of irreducible natural phenomena. It is not clear whether such a project can be carried out successfully, and it is dubious that it can be carried out at all without engaging in closer dialogue with hands-on philosophy of science. I regard this as a crucial research agenda for those wishing to preserve the dualist option. Likewise, the kind of comparison of dualism and various forms of materialism undertaken by Robinson is well taken, but it needs to be expanded into a conversation in which dualism is compared with various forms of pluralism as well.

6

Nonreductive Physicalism and Mysterianism

Of the traditional views in philosophy of mind, nonreductive materialism would appear, at least on first examination, to be the best able to accommodate scientific pluralism. To be sure, some nonreductive materialists, such as Davidson, have in fact held that mental phenomena (and things like linguistic and social phenomena that depend on them or are inextricably interwoven with them) are unique in their irreducibility. But there is nothing in either the nonreductionism or the materialism of nonreductive materialism that is incompatible with scientific pluralism. Many nonreductive materialists might be *surprised* to discover that it is "explanatory gaps all the way down," but it is not immediately clear that this should give them reason to lose sleep at night.

Additionally, nonreductive materialists already reject Negative EMC, as they hold that mental phenomena are both physical and irreducible. Therefore, unlike dualists, they do not have to worry about the pull Negative EMC exerts toward radical ontological pluralism; and unlike reductionists, they do not need to worry about the explanatory gaps being incompatible with a commitment to their breed of physicalism.

Indeed, the discovery of widespread explanatory gaps might even be seen as *bolstering* the case for nonreductive physicalism. One of the things that nonreductive physicalists have had to justify to their reductionist counterparts is how it is that the mental, and it alone, can supervene upon the physical without being reducible to it. But if irreducibility is widespread, this is no longer a problem. Nonreductive physicalists share with reductionists (and indeed with dualists) the assumption that chemical and biological facts are metaphysically supervenient upon fundamental physical facts, and this intuition is

unlikely to be shaken by the discovery of abiding explanatory gaps between these sciences. That is, in the wake of post-reductionist philosophy of science, we might seem to have reason to think that there are chemical and biological facts that supervene upon physical facts without being reducible to them. If this is the case, then it would seem to be a general principle that A can be metaphysically supervenient upon B even if it is not reducible to B. And if this is true in general, there is no reason to suppose that it cannot be true of mind-body relations as well. All in all, after the initial moment of surprise we all feel at the discovery of the disunity of the sciences, the nonreductive physicalist might well feel entitled to go home with a smile on her face.

6.1. Problems for Nonreductive Physicalism: The Nature and Evidential Status of Physicalism without Reductions

But I shall argue that all is not well for the nonreductive physicalist, and that scientific pluralism poses problems for her as well. Two of these, in particular, stand out. One concerns the *nature* of the "physicalism" or "materialism" the nonreductionist is to embrace. The other concerns the *evidential status* of physicalism and materialism once reductionism within the natural sciences has been abandoned.

The first problem, already alluded to in chapter 4, is that we need to clarify just what doctrine "physicalism" is to be if it is not the doctrine that all phenomena are *reducible to* physical phenomena. Is physicalism a claim merely about *objects*—for example, that all objects are physical objects? (If so, this would seem to allow dual-aspect theorists and perhaps even property dualists to count as physicalists.) Or is it a claim about both objects and *properties*—for example, that all objects are physical objects and all properties are physical properties? This would avoid the worry about property dualism counting as physicalism. But if "physical property" is to be a useful term at all, we will need some account of what it is for a mental property to "be" a physical property without being reducible to it. Or is physicalism a claim about *facts* about the world—that is, that the complete set of fundamental physical facts about the world determines all of the facts *simpliciter*? But what kind of "determination" are we talking about here? If it is *causal* determination, the thesis is compatible with an epiphenomenalist dualism. Likewise, if it is a purely logical form of determination—a kind of material conditional that accurately reports empirical relations between physical and mental facts (e.g., that a human experiences pain if and only if C-fibers are firing)—this is compatible with several forms of dualism. And unlike the causal account, this account is not explanatory. To see what a given nonreductive physicalism really amounts to, we will need to look at specific proposals.

A second, and more direct, problem arises from the role supposed reductions have traditionally played in motivating a case for physicalism. As long as one supposes that the sciences have produced widespread intertheoretic reductions, one might think that it has been *demonstrated* that at least large portions of the world around us are nothing but complicated physical processes. If a

natural phenomenon N is derivable from a (true) physical theory P, then we can *show* that nothing about N requires us to posit anything beyond P-phenomena. But if we can derive only *some* N-phenomena from P-phenomena, or none at all, we can make no such demonstrative claim, and indeed the situation invites the question of whether we may not need to suppose that there are aspects of N over and above what is necessitated by the P-facts.

Moreover, physicalism has historically been driven by the assumption that the sciences are steadily churning out intertheoretic reductions, or at least have begun to do so and can be expected to do so in the future. But if we adopt scientific pluralism—not only as a claim about the current state of science, but as a claim about what the sciences are likely to produce in the future—we no longer have such an inductive base on which to argue for physicalism. If one traditionally climbed up to physicalism via the ladder of reductionism, can one still reasonably be a physicalist once the ladder is kicked away?

In addition, the nonreductive physicalist owes us an explanation of how it is that relations can be metaphysically necessary but abidingly epistemically opaque. Traditional explorations of necessity linked it closely with the principle of noncontradiction: if P is necessary, then it should be possible to derive a contradiction from not-P. All necessities were conceived as epistemically transparent, at least to an ideal mind. But the nonreductive physicalist holds that they can be epistemically opaque, at least to minds like ours. We deserve an account of how this can be so.

In this chapter, I explore several forms of nonreductive physicalism that seek, in different ways, to address these issues. I argue that none of them is entirely satisfactory, and that nonreductive physicalism, though not inconsistent, must be regarded largely as a standpoint of faith.

(A note: it is debatable whether some of the strategies examined here—for example, various forms of identity claims—are better viewed as forms of reductive or nonreductive physicalism. I shall simply take the view that the hallmark of *reductive* explanation is its epistemic transparency, and hence that views that involve epistemically opaque identity claims—that is, that all A-instances are B-instances, or that this A-instance is a B-instance, without providing a way of seeing a necessary A-B connection—should be counted as nonreductive accounts. The taxonomy might turn out differently, of course, if one employed a different notion of 'reduction'. In any case, the objections to these apply regardless of where they are classified.)

6.2. Davidson, Anomalous Monism, and Interpretivism

Donald Davidson is the source of arguments that are often credited with having sparked the current popularity of nonreductive physicalism. Some of Davidson's views, however, are minority views even among nonreductive physicalists. And so I shall first explore the constellation of views representing his own position as best I can, and then address some of the constituent ideas divorced from their original context.

Davidson's metaphysical view is called *anomalous monism*. The monism consists in the claim that all entities and events are physical entities and events. The "anomalism" consists in the claim that there are no "psycho-physical laws" (in the sense of laws binding physical and mental states) such that a physical description of a human body and its behavior, or even of the entire physical universe, entails a unique assignment of mental properties. Davidson's claim here is *not* simply about an epistemic gap. He does not mean that there *is* a single right answer to what a person believes and desires, but an outside observer (and perhaps even the person herself) cannot be sure what that right answer might be solely on the basis of the physical facts. Rather, he claims that, metaphysically, the complete state of the physical world does not determine a unique assignment of such mental properties.

The *reason* given for the anomalous character of the mental derives from Davidson's interpretivist view of the nature of mental states. Ascriptions of mental states, according to Davidson, do not pick out intrinsic or internal states of a person, the way ascriptions of charge or mass do. Rather, they involve acts of interpretation: of ascribing a set of inferential commitments to the person that are consistent with his behavior, interpreted as being rational. (The assumption of rationality is Davidson's "Principle of Charity.") Davidson is thus advocating a view something like Dennett's (1987) claim that, in ascribing intentional properties, we are adopting a particular interpretive stance: the "intentional stance." However, it is *also* part of Davidson's view that there is always *more than one* consistent interpretation of a person's behavior that meets this criterion. Any set of behaviors—or at least any set that can be interpreted as rational at all—is compatible with multiple interpretations in terms of assignments of beliefs and desires. Hence any set of physical facts about a person's physiology and behavior underdetermines an interpretation in intentional terms. And since facts about mental states are *constitutively* interpretive—there is no question of "getting it right" among equally consistent interpretations—the physical facts do not determine a unique mentalistic interpretation because there *is no* fact about what mental state a person is in beyond the facts about how he may consistently be interpreted using the Principle of Charity.

Our ways of understanding other kinds of events, including those described in the physical sciences, do not have this complication: in understanding a nonmental event, we do not attempt to interpret it in terms of a rational set of beliefs and desires, and so no corresponding indeterminacy arises. (Of course, linguistic and social events are interwoven with the mental in such a fashion that the same issues apply to them, but it is all of a single piece, handled in the same interpretivist terms.)

Davidson is to be commended in having ponied up and given a reason for why he thinks mental events are not determined by physical events. Note, however, that he has done so in a fashion that bars us from applying at least one familiar definition of 'physicalism' to his view. His interpretivist view entails, not only that mental states are not reducible to physical states, but also that they are not supervenient upon them either. Or, at least, there is not a

single mental state interpretation, as opposed to a family of incompatible interpretations, that is supervenient upon the physical facts. His view is monistic solely in claiming that every token mental event is also a token physical event. Indeed, it is an *anomalous* monism precisely in holding that there are no necessary, or even nomic, mind-body connections.

One might sensibly ask at this point, "So what is physicalist or materialist about this view? Is it anything beyond a bald-faced *assertion* of token identity—and without supervenience at that?" I think such a question is in order. However, to do justice to Davidson, we might think of anomalous monism as representing a kind of *consistency proof* between monism and nonreductionism. Davidson's interpretivism shows us one way that the mental properties of a physical system might not be reducible to (or even determined by) its physical properties without falsifying monism: namely, if mental properties are not intrinsic properties, but properties that are a product of rational interpretation from the intentional stance. This interpretivist assumption generates the result that mental properties are underdetermined by the physical facts, or at least that the physical facts do not entail a single and canonical "correct" mentalistic description.

Davidson's strategy of treating ascriptions of mental properties as fundamentally different from ascriptions of other types of properties, because they involve a rationality-imputing, intentional-stance interpretation, has gained some adherents in philosophy of mind, perhaps most notably Daniel Dennett. However, it is not clear that even most materialists are convinced that it is the right way to understand mental properties. Many materialists, even in philosophy of mind, are deep realists, wanting to hold that mental properties are identical with, or emergent from, or resultant from, properties of the brain, or perhaps the brain in its interactions with its physical environment or in conjunction with facts about the selection history of brains in that biological lineage. And so, if Davidson's compatibility proof were the *only* way one could embrace both materialism and nonreductivism, even many physicalists would not be satisfied.

But even Davidson's strategy is challenged by scientific pluralism. Davidson more or less assumed that reductionism was tenable in the physical sciences when he put forward his anomalous monism, and viewed the mind as being unique in its irreducibility, because only it (and interrelated linguistic and social phenomena) requires rational interpretation. But if irreducible phenomena abound, then the mind is *not* unique in this respect. If one is independently drawn to Davidson's interpretationist account of mental state ascriptions, one might be prepared to say that these explain a *special* sort of irreducibility *in that one case*, leaving open why chemical and biological phenomena are irreducible. But if one is not independently drawn to it, one might think that a more unified explanation is desirable.

Of course, a kindred explanation of scientific disunity *may* be ready to hand from within the resources of the Quine-Davidson tradition. For within that tradition one also finds Quine's view that, even for the physical world, there are always multiple global "conceptual schemes" that are consistent with the behavioral facts about a person's description of the world, and indeed of the

world itself. Some aspects of Quine's view, such as that we do not know what we ourselves mean by 'rabbit'—whether it refers to rabbits or undetached rabbit-parts or temporal slices of rabbit-time-ribbons—have always struck me as ludicrous. (I don't know about Quine, but I know which of these I do and do not mean. There are no doubt indeterminacies in my usage, but this is not one of them.) But, those worries aside, Quine is probably correct in saying that there are always (for some sort of mind or another, and perhaps for human minds) multiple ways of dividing up the world conceptually.

Does this present a way of accommodating scientific pluralism? It *would* do so, if the scientific pluralism in question were in the form of *different and competing global conceptualizations* of the world (for Quine's conceptual schemes are global in nature). Or, if we liberated the view a bit from Quine's own holism, it might be a useful account if there were incommensurable but equally powerful theories *of the same phenomena*. And there are scattered instances in which this is the case. For example, for any set of data points, there are always multiple mathematical functions that fit them equally well. (Compare Kripke's [1982] claim that for any set of moves in a game, there are always multiple and incompatible rules that fit them, though they diverge with respect to predictions of future behavior.) And, to a more limited extent, there have been times when there have been incommensurable theories in science that seemed equally well empirically grounded, even if they did not always explain the same data: for example, particle and wave accounts in classical optics.

But the plurality of scientific models is in fact very different from the plurality of Quinean conceptual schemes. First, the models are not comprehensive and holistic, but more local: accounts of gravitation, or the Krebs cycle. Moreover, the problem is not that we have incommensurable, but equally good, accounts of the *same* phenomena, but that we have individual accounts of diverse phenomena that we cannot completely fit together into something like a single axiomatic system.

Indeed, far worse, an abiding scientific pluralism is incompatible with Quinean holism. If our knowledge is segmented, not globally, but in context- or discipline-specific chunks, then it is not clear that there can *be* anything on the order of an integrated and consistent "conceptual scheme" or "worldview." And while Davidson rejected Quine's notion of "conceptual schemes," he is equally committed to semantic and epistemic holism. As a result, it is not clear that he could accommodate abiding scientific pluralism without a drastic revision of some of his deeply held views.

6.3. Token Physicalism without Interpretivism or Holism

This leads us to a second strategy for nonreductive physicalists: the adoption of token physicalism without reductionism, interpretivism, or holism. Over the past generation, a great many philosophers have espoused views called "token physicalism." However, this label is actually applied to a number of views that are, for present purposes, importantly different from one another.

Token physicalism came into vogue in the 1970s as a result of functionalist critiques of type–type identity theory. Functionalists view mental state types (such as "pain") as functional types, and point out that they, like other functional types, are multiply realizable. Assuming that "pain" is a functional-kind term, humans, earthworms, Martians, and even robots might have functional states corresponding to pain, yet be built out of very different types of components, so that the "realizer" of the pain-function in each is unique. Nonetheless, human pain, earthworm pain, Martian pain, and even robotic pain could each be physical processes, in the sense that in each case the pain-function is realized entirely by some physical system or other. More radically, each instance of my having an occurrent belief with a given content (say, each time I think "Mark Twain wrote *Tom Sawyer*") might be realized by a slightly different pattern of neural activity in my brain; but each occurrence that realizes that belief-that-p function is *some* physical state or other. We may call this thesis—that each particular object and event is (perhaps among other things) a physical object or event—the *token identity thesis* or *token materialism*.

The word 'physicalism' is indeed sometimes used in such a fashion as to be synonymous with 'materialism'—that is, an ontological thesis about the inventory of the world. However, recent philosophical usage has tended to distinguish the two in subtle but important ways. First, 'physicalism' is often tied to the science of physics in a fashion that is agnostic about what sorts of notions will prove fundamental in the future of that field. If physics decides that it is force or energy that is fundamental, and not matter, then this might constitute a rejection of *materialism*, narrowly conceived. "Physicalism," however, can accommodate this possibility in advance by linking the notion of "the physical" to whatever objects, properties, and events physics might settle upon as needed for its domain. Second, the term 'physicalism' is generally understood to imply certain types of *determination relations* in addition to its assumptions about ontological inventory:

- *"Vertical" or "compositive" determination (or supervenience)*: at a time *t*, the sum total of the *physical* facts at *t* determines all facts *simpliciter* about objects, events, and properties at *t*. (Or, to accommodate historically rooted facts, all physical facts up to and including time *t* determine all facts *simpliciter* about objects, events, and properties at *t*.)
- *"Horizontal" or "temporal" determination (or "causal closure")*: every event at *t* has a completely adequate set of physical causes at times prior to *t*.[1]

The *token identity thesis* or token *materialism*, however, is itself neutral on both these points. It asserts merely that each bearer of mental properties is also a physical object. The claim that any object O has mentalistic properties M is also a physical object with some set of physical properties P is insufficient to entail (a) that the physical properties determine the mentalistic properties, (b) that M has an adequate causal basis in physical facts, or (c) that effects caused (in part) by O have a complete set of causes that need not involve M. *Token*

physicalism should thus be viewed as a stronger thesis than the token identity thesis/token materialism.

6.3.1. Token Materialism and the Identity Thesis

Token materialism, while incompatible with substance dualism, is compatible with property dualism, so long as it is a dualism that denies that there are any objects that have *only* mental properties and no physical properties. As a case in point, take Strawson's (1959) account of individuals. Strawson claims that *persons* (normal adult humans being a paradigm case, but not necessarily the only such case) are things to which both mental and physical predicates apply. (Something biologically human that was incapable of mental states—say, an anencephalic child—would not, I suppose, count as a person on this view.) One might adumbrate upon Strawson's view to describe a kind of person to which a token identity account of *events* applies as well: one in which every event that can be described in mental terms also has some physical description or other.[2]

But this does not require that the mental properties be metaphysically supervenient upon the physical properties, but only that, whenever some mental state ascription is true, *some physical description or other* of that event is also true. More important, it does not require that, once one particular physical description has been pinned down, a unique mental description is entailed. Nor does it require that there can be no changes in mental properties without changes in physical properties.

Moreover, token identity, as a thesis only about *objects* (and perhaps token *events*),[3] and not properties, does not restrict causation to physical causation. Someone taking a Strawsonian line on the nature of "persons," for example, might feel free to treat both sets of properties as capable of making independent causal contributions, perhaps on the model of gravitational and electromagnetic forces. Gravity and electromagnetism mark out not only distinct sets of *descriptions* and *properties*, but indeed two *independent* properties, each of which is capable of making causal contributions.

Token identity/token materialism in itself, therefore, is insufficient to ground any sort of supervenience thesis. One might be suspicious of the viability of a view that included token materialism but not supervenience and causal closure, but in terms of analysis, at least, those additional views are needed to get what is needed for physicalism.[4]

6.3.2. Token Identity Plus Supervenience

A first way to strengthen token identity in the direction of token physicalism is to add to it theses about "vertical" determination or metaphysical supervenience. That is, to hold that, for any object, state, or event O with mental properties M, O not only also has a description P as a physical object, state, or event, but that its having mentalistic properties M is metaphysically supervenient either (a) upon (local or intrinsic) properties P, or else (b) upon P plus some further set of physical properties consisting in its history and/or relations

to the environment. (The second option allows semantic externalism to be consistent with token physicalism.) Such a position yields a *nonreductive* thesis if one additionally denies that M can be derived from P (or the wider set of physical facts) through a reductive explanation. Such a position is sometimes called "token physicalism" even without the additional thesis of causal closure being specified.

This sort of token nonreductive physicalism may be consistent with scientific pluralism, but it is simply the *assertion* of a position that does nothing to further the *plausibility* of the supervenience claim. To the extent that the prior plausibility of physicalism was underwritten by the assumption of the possibility of broad reductions, physicalism is still in need of a new evidential base in order to count as more than a standpoint of faith.

6.3.3. Causal Closure

Unlike the claim of supervenience, the claim for the causal closure of the world under physics *does* provide an independent reason to embrace physicalism. Indeed, a principle of causal closure has in recent years played an increasingly prominent role in arguments for physicalism.

The thesis of causal closure (TCC) is not directly a claim about supervenience, nor indeed about any other interlevel relationship; rather, it is a claim about the relations between prior and posterior events. As David Papineau (2001, 8) phrases it, it is the thesis that "all physical effects are fully determined by law by prior physical occurrences." Thus put, TCC is, strictly speaking, compatible with substance and property dualism and with Davidsonian indeterminacy of the content of intentional states. However, TCC has been used to paint dualists, at least, into a corner. If all physical effects are determined by physical laws and prior physical occurrences, then mental events play no independent causal role in determining any physical events. If TCC is true, one can assert dualism only at the cost of epiphenomenalism—the view that mental states play no causal role. (Or at least no causal role in determining physical events, as is required for an explanation of actions in terms of prior mental states. TCC, as enunciated by Papineau, is compatible with a completely independent causal chain of mental-mental causation.) While some dualists have been prepared to accept epiphenomenalism, it is a bitter pill to swallow. On the one hand, it precludes libertarian free will and agent causation. On the other hand, it is a blow to the ontological credentials of mental states if these depend on those states entering into causal relationships, or being needed as theoretical posits of a causal theory. If TCC is true, then mental states can enter into causal relations only by dint of their identity with physical states.

While TCC is indeed a popular doctrine today and is widely employed as a premise in debates about mental causation and the philosophy of mind, it is a premise that has been adopted largely without argument or scrutiny. This fact has indeed been noted by one of its more important advocates, David Papineau. In *Thinking about Consciousness*, Papineau (2002, 45) indicates that he had originally thought that the thesis of causal closure under physical laws was not a problematic issue:

> The one assumption that I did not expect to be uncontroversial was the completeness of physics. To my surprise, I discovered that a number of my philosophical colleagues did not agree. They didn't see why some physical occurrences, in our brains perhaps, shouldn't have irreducibly conscious causes.
>
> My first reaction to this suggestion was that it betrayed an insufficient understanding of modern physics. Surely, I felt, the completeness premise is simply part of standard physical theory. However, when my objectors pressed me, not unreasonably, to show them where the completeness of physics is written down in the physics textbooks, I found myself in some embarrassment. Once I was forced to defend it, I realized that the completeness of physics is by no means self-evident. Indeed, further research has led me to realize that, far from being self-evident, it is an issue on which the post-Galilean scientific tradition has changed its mind several times.

We would do well to take Papineau's lesson to heart. Indeed, while Papineau's admissions on this subject reflect admirable intellectual honesty, his own attempts to address the issue in the appendix to *Thinking about Consciousness* are curiously unsatisfying. There, he observes that, by about 1900, there were two areas of scientific inquiry in which there was serious question about whether it was necessary to include nonphysical causal principles: living systems and the conscious mind. He then provides a brief overview of developments in the life sciences over the course of the twentieth century that unlocked a number of mechanisms underlying processes like metabolism, and thereby stripped vitalism of much of its previous allure. But—and this is a curious fact given that the book in question is a book about consciousness—he never addresses, in similar fashion, how far neuroscience might go in addressing cognate concerns about the mind. Instead, he seems to argue by *analogy*: that because developments in the life sciences have progressively provided alternatives to the vitalist impulse, it is reasonable to expect that future work in neuroscience or other disciplines will do the same with respect to the antiphysicalist impulse.

Such an argument is unsatisfactory in a number of regards. Most immediately, it does nothing to push forward the conversation with contemporary dualists and other antiphysicalists, none of whom are inclined to be vitalists. Chalmers, for example, would see biological properties as functional properties, and hence candidates for mechanistic explanation and also supervenience upon the physical, but he does not view consciousness as a functional property that is subject to this form of relationship to a "realizing" system. Contemporary dualists like Chalmers already have a reason to think that consciousness is *dis*analogous to biological properties in crucial ways that block the argument by analogy in a principled fashion.

Going a bit deeper, there are a number of ways in which the examination of scientific explanation over recent years might call Papineau's analysis into question. For example, does his position require the kind of picture of a unified

science—or even a unified physics—that has been assaulted by recent philosophy of science? If so, it is imperiled by the evidence for the disunity of the sciences canvassed in chapter 3. However, Papineau does not need to endorse a unity-of-science claim. TCC, as asserted by Papineau, does not require that there be a single, unified physical theory that can explain all physical effects, but merely that, among the many theories, laws, and models that are used to describe physical systems, there is always at least one (or some combination) of these to explain each token event. And so TCC is available to support a nonreductive form of physicalism.

However, recent philosophy of science's repudiations of the Positivist/Empiricist view of science and explanation are not confined to the rejection of Carnap/Nagel-style reductionism. Another crucial turn in philosophy of science has been the recognition that each physical theory, law, or model is *idealized*—sometimes in ways that render it incommensurable with other physical theories, laws, and models (Cartwright 1999; M. Wilson 2006). This, in turn, has implications for how we ought to view claims about explanation, determination, and supervenience. If the bulk of physical explanations are in fact only partial explanations, are idealized, and are counted as "adequate" explanations in a given instance on pragmatic grounds, the pluralist who is inclined to embrace Negative EMC (the thesis that failures of reduction imply failures of metaphysical supervenience) might well be inclined to accept the science but see its proper analysis as challenging rather than supporting *metaphysical interpretations* of the science that view science as committed to causal determinism or metaphysical supervenience. More generally, Papineau's approach still betrays an assumption common to many advocates of TCC: the assumption that notions like causal determination and supervenience can be read off the first-order commitments of the sciences themselves, rather than involving additional and substantive metaphysical theses.

6.3.3.1. Causal Determination. Consider first the question of causal determination. Does the availability of a theory, law, or model that explains an effect entail that the effect was fully determined by the causal factors alluded to within the theory, law, or model? I think the answer to this is *no*, and for reasons having nothing to do with the philosophy of mind.

Consider a simple textbook-style physical explanation, say, of the motion of a falling object close to the Earth using the gravitation law. A gravitational model of falling objects *does* contain substantive commitments to real gravitational invariants that play a causal role in real-world ballistic situations. But it does not involve a commitment to the further notion that *only* gravitational invariants are in play in a given case, even a case where a gravitational model is good enough for prediction and explanation. Other factors, such as electromagnetism and aerodynamics, may also be in play even in the simplest cases, but are sufficiently small in their contribution to the resultant kinematics that employing a purely gravitational model is good enough for practical purposes. But in other cases—say, when the object dropped is a paper airplane rather than a tightly wadded ball, or when the object is metallic and a magnet is

present—these other causal factors may matter a good deal to the resultant behavior, and a gravitational model alone is inadequate to the tasks of explanation, description, and prediction. Indeed, even in the cases in which a gravitational model alone is good enough for practical purposes, the results of plugging initial conditions into that model never predict real-world behavior with complete accuracy, as other forces are always in play. At best, real-world behavior is determined by a *complete summation of forces*, iterated over real time, over the initial conditions. But sometimes there is *not* a technique for summing forces (Cartwright 1989, 1999). And the disparate physical models we employ to explain different parts of a phenomenon are sometimes idealized in incompatible ways; for example, as Mark Wilson (2006) points out, crucial terms like 'force' may have divergent and incompatible meanings within the context of different models. (Such issues are treated further in chapter 7.) Additionally, where quantum randomness is potentially a factor (which may be all the time), the outcome is underdetermined by even the combination of laws and initial conditions.

So, on the one hand, the models we actually employ, usually piecemeal, to describe or explain a physical phenomenon tend to underdetermine the real-world behavior. And, on the other hand, the idea that resultant behavior is determined by "a summation of *all* the forces that are in play," which looks so attractive in the abstract, often proves unworkable or even unintelligible when one turns to the details of the models employed by the scientist or the engineer. This forces us to reexamine what we mean by such claims as "All physical effects are fully determined by law by prior physical occurrences." If this is a claim about the kinds of explanations that are actually used by scientists and engineers, then the claim is simply *false*. Single models seldom if ever fully account for the complexity of real-world kinematics. And the kinds of examples surveyed by Cartwright and Wilson show that it is often impossible to perform a true summation of forces.

It is important to emphasize that the problem here is *not* simply the *computational intractability* of most such summations—the kind of problem one encounters in systems so simple as three bodies with only gravity in play. Rather, the problem is that the ways that real scientific models are idealized often renders them unsuitable for integration with one another by way of vector algebra. Nor does there seem to be any reason to believe that this is simply a symptom of the present immaturity of our scientific modeling or the intellectual laxity of scientists and engineers. Rather, it seems to indicate a systematic and principled mismatch between the types of models actually employed in the sciences and what would be required for complete determination of real-world events.

The alternative here would seem to be to try to separate our notions of "physical causation" and "determination" from the types of explanation and modeling actually found in the sciences. Here, I think, nonreductive physicalists ought to feel two countervailing instincts. On the one hand, the movement to distinguish "physicalism" from "materialism" by a closer connection with real physics pulls in the direction of keeping notions like "causation" and

"determination" anchored in actual practices of physical explanation and modeling. On the other hand, having already allowed that there can be "vertical" determination relations (supervenience) that are not underwritten by reductive explanations, the nonreductive physicalist might find it easy to countenance a similar possibility for "horizontal" or "temporal" (causal) determination relations that might outstrip what can be derived from physical laws, theories, and models. There might *be* aggregate causal factors that jointly determine physical events, even if our ways of modeling these are necessarily imperfect, idealized, and piecemeal. Such a move, however, comes with a high price tag, especially for a would-be naturalist. It moves claims of "causal closure" and "causal determination" out of the realm of things that can be read off the sciences themselves, and into the realm of metaphysical speculation. In so doing, it runs the risk of accounts of causation and determination becoming a free-spinning wheel, disconnected from the driveshaft of real science.

These considerations run counter to a widespread assumption that a commitment to scientific laws implies a commitment to determinism. Why is this assumption so widespread? One reason might lie in the influence still exerted by the Positivist conception that laws are universally quantified claims about the real-world behavior of objects. On this view, for example, the gravitation law makes claims about how real objects always fall. But such an interpretation of laws is problematic. If this were what the gravitation law said, that law would be patently false, as the examples of the paper airplane and the metallic object dropped near a suitably strong magnet show (compare Cartwright 1983, 1989, 1999). For this reason, philosophers of science have largely abandoned the Positivist conception of laws in favor of the view that laws like the gravitation law express "causal powers," or as I prefer to phrase it, "potential partial contributions to real-world kinematics" (Horst 2004). On the causal account, laws do not individually state how things in fact behave, but rather express individual causal invariants that play a role in real-world kinematics. Each individual law is abstract and idealized, in that it brackets off other causal contributors to isolate a single set of invariants. (Many additionally make "distorting" idealizations, such as treating objects as point-masses or collisions as elastic. See discussion in chapter 7.) Hence commitment to the gravitation law does not involve a commitment to the further thesis that any real ballistic situation is determined by that law plus initial conditions, because a commitment to the gravitation law leaves open-ended the question of what other causal factors, such as aerodynamics and electromagnetism, may be in play.

Papineau would, I think, be happy to allow this. He can take the position, after all, that real-world behavior is determined, not by any single law, taken in isolation, but only by the combination of all of the nomic forces acting upon the situation. Indeed, Papineau goes so far as to say that there is no inconsistency between very general physical laws, such as conservation laws, and the possibility of independent vitalistic or mentalistic *laws*. What he thinks is ruled out is *anomic* causation (see Papineau 2002, 248–49). But this, too, involves an assumption that goes beyond the sciences themselves. The latter involve the enumeration of laws and mechanistic models that aim at revealing real causal

invariants. But they do not themselves involve the claim that *all* causal factors must be nomic in character. Neither true randomness nor agent causation is precluded by a commitment to laws, any more than commitment to one law precludes commitment to another.

It is important to distinguish several types of claims here that are easily conflated. One is a *methodological maxim: look for* laws and for causal mechanisms that are nomic in character and appeal only to physical phenomena. This may be a useful maxim in guiding scientific research, but it should be differentiated from two other claims. A second claim is that the scope of science is restricted to phenomena that are nomic. This is a more controversial claim, and it is not a claim *within* the sciences, but a philosophical claim *about* them. (It is controversial, in part, because of the role that notions of "randomness" may play in postclassical physics, and the question of whether probabilistic equations are truly "nomic" even though they are probabilistic rather than deterministic. It is additionally problematic because a number of sciences have many models without laws.) But even this second claim must be differentiated from a third, metaphysical claim: that *all causal factors must be nomic*.

This third claim is not entailed by our first-order commitment to the laws and models produced by the sciences themselves. It is compatible with the truth of the laws of gravitation and strong/weak/electromagnetic force that there might also be anomic causal factors, as well as true randomness. And even if one holds (on transempirical grounds) that any such factors fall out of the scope of *science* (if that is understood as restricted in scope to nomic phenomena), this is still compatible with there *being* anomic causes. That is, one must distinguish the following theses:

(C1): Laws $L_1, \ldots, L_n$ are true.
(C2): $L_1, \ldots, L_n$ are all the laws there are.
(C3): Events are causally closed under $L_1, \ldots, L_n$.

Causal closure is established only in C3, and C3 is not entailed by either C1 or C2. It is an additional principle, one that must be added to a commitment to laws, and even to the philosophical premise that the domain of science is restricted to nomic phenomena. The upshot of this is that, if we adopt a causal understanding of laws, a commitment to the truth of those laws does not imply a commitment to determinism and is compatible both with true randomness and with anomic causation, including free-agent causation.

Could a principle to the effect that all causation (at least of physical effects) is fully determined by physical laws be established empirically, through the sciences, rather than as a matter of philosophical taste? My own views on this question are pessimistic. There are two basic ways to approach the question empirically, one direct and one indirect. The *direct* way would be to perform experiments, say, on mind and brain, that investigate whether there are effects found there that resist explanation in terms of physical laws. The *indirect* way would be to find a law of physics that is truly incompatible with indeterminism or agent causation (i.e., a principle found in first-order scientific claims, rather

than in a philosophical interpretation of the sciences). Papineau and others suggest that such a principle can be found in the conservation laws, and this is a useful example to use in exploring both the direct and the indirect strategies. Of the conservation law Papineau (2002, 249) writes:

> The content of the principle of the conservation of energy is that losses of kinetic energy are compensated by buildups of potential energy, and vice versa. But we couldn't really speak of a 'buildup' or 'loss' in the potential energy associated with a force, if there were no force law governing the deployment of that force. So the very idea of potential energy commits us to a law which governs how the relevant force will cause accelerations in the future.

That is, the principle of conservation is (a) general in scope, and (b) nomic in character. Because it is general, it ranges over all physical effects, and because it is nomic, this excludes the possibility of anomic causation (mental or otherwise) of physical effects.

It is true that the conservation principle is understood to be "general in scope" in the sense that it is not a principle that is applied only to a restricted class of physical interactions, unlike, say, models that are applicable only to cases of laminar flow or to matter in a particular phase. But does this mean that the conservation law can be properly interpreted as a universally quantified claim about all physical events? Not necessarily. Often, laws that were once thought to have general scope have later been found to apply only to a limited (though significant and perhaps large) range of conditions. When such a discovery is made, we do not reject the law, but merely revise our understanding of its scope. The core of the conservation principle lies in the claim that, under some broad range of conditions, there is a quantitative trade-off between kinetic and potential energy. This principle is well-verified for a wide range of cases in mechanics and thermodynamics. Indeed, unlike many other physical laws, there are no laboratory cases where it is known to break down.

This, however, does not entail that it is truly universal in scope, nor that it may be safely imported to what may be very different situations, such as human action. If one accepts the premise that anomic causation (and hence anomic *agent* causation) would violate the principle of conservation, then either (a) there is no anomic (agent) causation, or else (b) the principle breaks down in cases where anomic causation is brought to bear. That the principle would break down in such cases may seem unlikely, but it is not clear that it seems more unlikely than did many other assumptions about classical mechanics whose aptness has proven to have limited scope. I would not go so far as Nancy Cartwright, who occasionally claims that experiments in carefully controlled laboratory environments give us literally no reason to expect that things will behave in the same ways outside of those environments. But it does seem right to be cautious about assumptions concerning scope on the basis of constrained laboratory experiments.

The question of whether there are cases of anomic causation where the principle of conservation breaks down is thus, in some sense, an empirical

question. Unfortunately, I fear that it is an empirical question that may prove intractable, at least in the case of the mind. Suppose, for example, that we wished to put matters to the test, and determine directly whether cases where we have *prima facie* reason to believe that there is agent causation involve violations of the principle of conservation of energy. To do this, we would certainly need to take very exacting measurements of energy within the brains of living conscious subjects over the time frames involved in decision-making and action, as well as taking careful stock of the neural and subneural processes (e.g., metabolic processes) going on that might contribute in a known physical fashion to thermodynamic changes. These tasks in themselves very likely far exceed present experimental methods, and may in fact be impossible to perform noninvasively. We would then need to compare these to the values we would *expect* to find in the absence of anomic mental causes. Here, we do not know what size of effect we should be looking for—that is, we don't know how much of a difference in energy we should *expect* if there should be anomic agent causation—and hence there is a fundamental problem for experimental design in distinguishing differences in uncontrolled variables from innocuous variations within a margin of error. In addition to these problems, it is not clear that we should *expect* anomic agent causation to add energy to the system (a concern that also affects the indirect argument). While there is no doubt an empirical *fact* about each of these questions, it is not clear that we possess, or realistically might ever possess, empirical means of determining those facts.

In short, *indirect* proofs from conservation are question-begging, as (a) anomic causation might be thermodynamically neutral, and (b) if it is not neutral, this might mark a restriction in the scope of the principle of conservation rather than evidence against anomic causation. Direct investigation, on the other hand, seems likely to remain empirically intractable.

6.3.3.2. "Vertical" Determination and Supervenience. The argument for causal closure also serves as the basis for an argument for psycho-physical supervenience: if TCC is true, then token mental states can cause physical effects only by dint of their identity with token physical states. (Conversely, if TCC is true, one can endorse independent mental states only at the cost of psycho-physical epiphenomenalism.) However, this argument is only as good as its premises, and TCC is employed in this argument as a premise. Given the problems already discussed for the evidential status of TCC, this argument is rendered suspect as well.

The relation between causal closure and supervenience is also deserving of some scrutiny. If, like Papineau, we locate evidence for causal closure at the level of *laws*, what we have, at best, is evidence for "nomic" supervenience. Laying aside general concerns about this notion, which are discussed elsewhere in this book, it is sufficient to stress that evidence in the form of lawlike relations falls short of establishing claims of *metaphysical* supervenience, which I am taking to be partially constitutive of physicalism. In the previous section, we were concerned with causal determination, and hence followed Papineau's lead in looking at the issue of anomic causation. But one can reject

physicalism without embracing anomic causation. For example, one can hold that mental states are nomically but not metaphysically supervenient upon physical states, while embracing either epiphenomenalism or a deterministic "downward causation" that is underdetermined by strictly physical laws. While I do not wish to commend either of these views to the reader, their compatibility with TCC further undercuts its use to argue for physicalism.

6.3.4. Token Identity and Causal Closure: Summary of Problems

The token identity thesis is not, by itself, sufficient to define a viable form of physicalism, as it is compatible with alternative views such as property dualism. Token physicalism helps define a consistent nonreductive physicalist position but provides no argumentative basis on which to prefer physicalism to its alternatives. The thesis of causal closure attempts to provide such an argumentative basis, but suffers from several problems. Some of these have to do with an implicit understanding of scientific laws and with conflations of commitments to first-order scientific claims with commitments to additional philosophical theses about the sciences. Others have to do with the fact that there are nonphysicalist positions that are compatible with causal closure.

6.4. Contingent Type-Identity

Another resource available to the nonreductive physicalist is contingent identity (or, perhaps better, contingent ident*ification*). To the extent that there are psycho-physical regularities, such as the co-occurrence of C-fiber firings and pains, these are subject to a number of metaphysical interpretations: interactionist and parallelist substance dualisms, occasionalism, property dualism, reductionism, necessary type-identity, and contingent type-identity. Even if necessary type-identity is ruled out on grounds of multiple realizability, it is still feasible to salvage a relatively strong relation in the form of contingent type-identity. The basic idea behind contingent identity might be put like this: while multiple realization has shown us that there is no metaphysically necessary biconditional relation between, say, C-fiber firings and pains, it has not shown that, for particular sorts of organisms (say, humans), such events may not, nonetheless, be type-identical. Martian pains may *be* flowings of green goo, and human pains *be* firings of C-fibers. (Take, as a paradigm, the intuition that we may say that temperature *is* mean kinetic energy in the case of gasses, even though this property identification will not serve in the case of solids or plasmas.)

This claim, however, seems to me to admit of two very different readings. One is that, in each case—human, earthworm, Martian, robotic—pain can be reductively explained by the relevant physical category in a fashion that is unproblematically epistemically transparent. The other is that, in each case, the type 'pain' is type-identical to some physical type, *even though the relation between being-a-pain and being-a-C-fiber-firing is abidingly epistemically opaque*.

The first interpretation is really an instance of broad reductionism, and hence I shall not consider it further here.

What, then, does the second interpretation really amount to? It seems to me that it amounts to the claim "Properties A and B are the same property, but we have no idea of how this might be so, and not only that, but nobody will ever have an idea of how it might be so." Arguably, not a very promising start. But we should at least consider some analogies. Suppose some visionary had predicted, circa 1300, that temperature (in gasses) is identical to properties of the motions of bits of matter making up the gas. (He could not have articulated this using the notion of mean kinetic energy, of course, as that notion was not yet available at the time.) One could not, at that time, have seen why this might be so. But later we found out how it might be so, and in fact *is* so, through the derivation of the gas laws from statistical mechanics. But this example has important differences: the fact that we can *derive* the gas laws means that the relation between *those* properties *is not* and *was not* in fact (abidingly) epistemically opaque. The nonreductive physicalist must hold, perforce, that this is *not* parallel with the case of psycho-physical identities. The problem is that, short of this sort of derivation, we are left with a *correlation* of variables, and such a correlation is susceptible to various metaphysical interpretations, such as causal covariation and identity.

That is not the only problem, of course. Another concerns the individuation of properties. One reasonable interpretation of properties is that properties are individuated by mode of presentation. On this view, temperature is not the same *property* as mean kinetic energy (even for gasses). Rather, temperature is *explained by* mean kinetic energy by way of a CAE, even if some other thermodynamic phenomena (such as entropy) may not be thus explained. However, an alternative approach is to view properties (or at least *some* properties) as natural kinds—that is, mind-independent real essences in nature, to which the mind might have multiple modes of access through different modes of presentation. On this interpretation, temperature and mean kinetic energy might be seen as multiple modes of presentation of a single property, which has no single canonical description.

This approach may initially seem tempting. But note that it is bought at a serious cost. It would seem to involve the invocation of something like the Lockean distinction between *qualities* or *nominal essences* (ways the mind represents things—more or less Fregean modes of presentation) and *real essences* (unknowable ways the world is in itself). This move is in some ways congenial to nonreductive materialism. Nonreductive materialists hold that at least some property relations are not fully intelligible. But the move to Lockean real essences requires more than this: it requires that the real and fundamental properties of things lie outside of human ken altogether, and that *all* our ways of conceiving things (including those of fundamental physics) are matters of how objects affect us, rather than of how they are in their own right. In fact, in part III, I endorse a variant of such a position. But I expect it is not really what most nonreductive physicalists are looking for.

Contingent identity is also notoriously problematic in light of Kripkean semantics and metaphysics, which claims that all identities are necessary.

If we hold that properties are independent of mode of presentation, then it seems that they ought to behave, semantically and metaphysically, like individual objects. But in that case, if properties A and B are identical, they are necessarily identical, because they are *the same property*, just as Sam Clemens and Mark Train are necessarily identical, because they are *the same person*. And so, if properties are individuated by mode of presentation, they are not identities, because a different mode of presentation implies distinct properties. And if they are individuated independently of mode of presentation, they are not contingent identities, because identities are necessary.

6.5. Occam's Razor

I have claimed at various points that robust mind-body correlations are compatible with multiple metaphysical interpretations: interactionist, parallelist, and occasionalist dualisms, reductionist and type-identity materialisms, interpretationism, and so on. In my experience, the standard response to this claim is that, while it is true that correlations are *compatible* with multiple metaphysical theories, this does not mean that all such interpretations are necessarily on an equal footing. Most conspicuously, reductionism has a kind of pride of place in being able to produce *demonstrations* of its preferred relation. But while the others do not enjoy this advantage, it does not mean that they are all created equal. In particular, there is this wonderful principle called Occam's Razor, which says that, given a choice between two equally explanatory hypotheses, one should prefer the simpler of the two. Identity theories and interactionist theories may enjoy equal explanatory power, but identity theory is the simpler of the two, ontologically speaking, since it does not postulate additional entities. Thus, invoking Occam's Razor, it wins out. Other things being equal, one substance, or one type of fundamental property, *is* better than two, at least for purposes of theory.

I suspect that St. William of Occam spends a great deal of time rolling in his grave. I confess that I have always been somewhat hesitant about Occam's Razor. For one thing, theoretical simplicity is only one of a large set of possible theoretical virtues and must be weighed against the others. Additionally, there are many ways to assess "simplicity." One theory may have fewer *entities* but more *laws* or *principles*, while another may multiply entities to gain simplicity in laws and principles. But perhaps most important, Occam's Razor is intended as a principle to apply only when adjudicating between two accounts that are indiscernible in terms of explanatory power. For example, here is an eminently simple theory: there is just one thing, a duck. Or, if you like, one lepton. Really simple ontological inventory. Very impoverished explanatory power. Alternatively, Spinoza's philosophy is wonderfully simple in its inventory: there is only one substance. But I do not see a rush toward Spinozism on that account. The point is that Occam's razor is, at best, applicable in deciding between two explanations with identically comprehensive explanatory power. It tells us nothing about what to do, for example, if theory A has a trifle more

explanatory power than theory B, but is slightly more complicated. Nor does it tell us how to measure explanatory power or complication.

Now this may work well enough when we are comparing two theories that are "of identically comprehensive explanatory power" in the sense that they explain the same range of phenomena, explain them in a satisfactory way, and explain them equally well. But this is not the situation we are in with respect to metaphysical interpretations of psycho-physical relations. These *are*, in some sense, "of identically comprehensive explanatory power," but only in the trivial sense that *none of them explains psycho-physical relations at all*. Reductionism, of course, is the exception to this: a reductive explanation *would* explain a lot. But without reductions, the remaining alternatives are all on an equal footing in terms of explanatory power, *but only because none of them explains anything at all*. Neither identity nor causal covariation is a theoretical relation that explains anything about properties. (That is, the *metaphysical* interpretation adds nothing beyond what is present in the science.) Occam's Razor may be of at least some use in adjudicating between theories that really could explain something. A theory that really explains A in terms of B is to be preferred over a theory that explains A in terms of B plus C but has no additional explanatory power. But it is not clear why a theory that *fails* to explain A in terms of B is to be preferred over a theory that fails to explain A in terms of B-C relations.

But perhaps this is a bit quick. It *does* seem that at least *some* identity claims have some explanatory power. For example, suppose we want to know why Sam Clemens was invited to be present at a signing of Mark Twain's *Tom Sawyer*. This is admirably and elegantly explained by the fact that *Sam Clemens is Mark Twain*. This, however, is an identity claim about an *individual*, not a *property*. But we might invoke examples involving properties as well. Why does the temperature of a gas increase with the increase of mean kinetic energy of the gas molecules? Because *temperature (in a gas) is mean kinetic energy*. But, again, this is a case in which the identification of notions from two scientific theories is underwritten by a derivation that is, or is at least very close to, a CAE. We do not simply identify temperature and mean kinetic energy on the grounds that doing so simplifies our ontology; we do so because we can derive a significant set of thermodynamic phenomena from statistical mechanics. We do indeed have an ontological simplification here, but the identification by which it is accomplished is posited only on the basis of powerful intertheoretic derivations. By contrast, we have *no* corresponding way of understanding basic phenomenological, intentional, or normative properties in physical or neurological terms. In the absence of such an explanatory relation between properties, we lack the very reasons we have in the case of the derivation of the gas laws for grounding a hypothesis of property-identity. The postulated identity between mean kinetic energy and temperature (in gasses) is grounded not so much in the fact that it simplifies ontology, as that it has real explanatory force in the form of mathematical derivations of properties. We have nothing of the sort in the case of correlations between phenomenological "feels" and brain states.

Indeed, one might, somewhat tentatively, take this line of reasoning even further. In the sciences, a nomic relationship between variables is often taken

to indicate that one is (partially) dependent on the other, but seldom is taken as indicating that they are identical. Pressure and temperature covary, but are not the same thing. To the extent that there is covariation between A and B *without* exhaustive explanation of one in terms of the other, we have reason to suspect that A and B are *not* identical.

It therefore seems to me that Occam's Razor is of little use to us here. I begrudgingly admit that there are very limited contexts in which it is of use. But mind-body relations are so far outside that context that I view the invocation of Occam's Razor as being, along with patriotism and consistency, one of the last refuges of a scoundrel.

6.6. New Semantics and Identity

A more powerful apparatus for dealing with epistemically opaque identity relations is supplied by the New Semantics developed by Kripke (1972) and Putnam (1975). NS was developed in opposition to traditional definitionalist semantic accounts, such as that of Mill. On the definitionalist account, the meaning of a term is determined by the sense of the term. There are multiple variations of this account, but to take one variant, the meaning of 'water' might be determined by the definition of 'water': for example, that it is a viscous, potable liquid often found in ponds, streams, and lakes. (On a slightly different view, Inferential Role Semantics, what is constitutive are the implications a speaker would draw: e.g., that she would infer, from the belief that something is water, that it is a liquid.) On this view, things similar to the senses (or whatever methods we use in identifying them), found in similar circumstances, in different worlds, would all count as water, regardless of their underlying nature.

Kripke and Putnam suggested, famously, that this is not how terms like 'water' in fact operate. In fact, if there is some Twin-Earth, a world whose macroscopic differences are indiscernible to ordinary perception from those found on Earth, but in which the viscous, potable liquid found in ponds, streams, and lakes is of a completely different molecular nature, say, XYZ, then the stuff they call "water" there is *not* among the reference class of our word 'water', even if there are people on Twin-Earth who look and sound just like us, use a word pronounced <WUH-ter> to refer to it, and use that word in all the contexts that we use the Earth-word 'water'. The reference of a term is not fixed (purely) by definition or inferential role. Rather, such words have the function of pointing to "natural kinds" that are involved in the right sorts of ways in the coining of terms.

Kripke's and Putnam's versions of this idea proceed slightly differently. Kripke suggests that there is, first of all, a "baptismal" context in which a word is anchored to its referent. We might, for example, on some occasion use 'water' to refer to a particular sort of stuff we see over there in the stream, which we identify as being clear, potable, frequently found in streams and lakes, and so on. But what 'water' refers to is *that* stuff, *whatever it is*, rather than *all* things that meet the identifying criteria of clarity, potability, and so on.

Likewise, it applies to instances of the same substance that lack the identifying characteristics, such as ice and muddy water. The reference is fixed by the baptismal context, which might be understood ostensively. After that, reference is *transmitted* causally: *I* mean by 'water' the stuff that the person I learned the word 'water' from referred to; she meant what the person she learned it from referred to, and so on back to the baptismal context. Kripke's theory is *ostensive/baptismal* in the original instance, and *causal* in *transmission*.

Putnam, by contrast, proposed a theory that is "causal" in a different way. Suppose that there is another world called Twin-Earth—not a *possible* world, but another *actual* world somewhere distant in space—which is exactly like the actual world, with the exception that wherever there is gold (i.e., the element Au) here, there is a compound EFG there. EFG is exactly like Au in all its obvious phenomenological properties, but different at the molecular level; for example, it is a compound rather than an atomic type. And indeed Twin-Earthers refer to EFG as "gold." (In Putnam's original example, it is water, not gold, that is twinned in this way. It is an obvious problem to suppose there are molecule-for-molecule duplicates of *us* in a world that is devoid of H_2O, as our bodies are largely composed of H_2O. Putnam is, of course, aware of this difficulty with the example, and presumably will not mind if I adapt it to involve an example that is not impossible.) On Putnam's account, the reference of 'gold' is determined, not by the baptismal context, but by the causal covariations between the lexical unit and its referent. When I say "gold," I refer to Au, because that is the stuff that I have been in causal contact with. When my twin says "gold," he refers to EFG, because that is what he has been in causal contact with. The story to be told about a triplet living in a world with *both* Au and EFG would be correspondingly more complicated.

Now the magic of both of these accounts is that they drive a wedge between the *sense* of the account (the identifying criteria and constitutive implications) and the *referent*. And, as a result, they at least *seem* to drive a wedge between conceptual entailment and metaphysical necessity. For example, suppose the conceptual content of my concept WATER is "clear, potable liquid found in lakes and streams." This content may be what I use to identify water, but it by no means entails that this stuff is H_2O. There are many possible worlds in which there are other substances (e.g., XYZ) that play this role. It is, by contrast, very likely the case that a physical-chemical description of H_2O would entail that it have the relevant phenomenological and commonsense properties, given other physicochemical suppositions and psycho-physical correlations. So perhaps what we ought to say is that the account shows that the sense of the term (particularly as used by someone who is scientifically uneducated) need not entail the referent, but does not show that an *adequate understanding* of the referent would not entail the sense. (In other cases, such as 'whale', the original sense might have involved constitutive implications such as "is a fish" that would prove to be false without this having the implication that the term was vacuous. And a more adequate understanding of whales might result in conceptual and linguistic changes that would eliminate the erroneous implications.)

Now the vital assumption in the argument about mental terms is that they behave like "natural-kind" terms such as 'water' and 'gold'. I happen to think this is quite dubious, for reasons I developed in chapter 2 and will return to anon. But first let us examine how this analysis is supposed to work. On this analysis, 'pain' might be characterized as follows: "*that* state, whatever it is, that is 'ouch-y.' " By dint of the ostensive ("that") element, this picks out something that is not "ouchiness" itself (the identifying criterion), but whatever makes the particular ostensive (or covariational) class "ouchy." Suppose that this is, in my case, but not in that of Marty the Martian or Robby the Robot, C-fiber firings. In that case, on the NS account, my use of 'pain' refers to C-fiber firings, just as my use of 'water' refers to H_2O, even though it is impossible for me to infer this from the sense of the terms. It is the property of *being-a-C-fiber-firing* that my term picks out, even if I do not know the nature of this property.

It is important to distinguish three levels at which such an account might be addressed. First, is such an account the correct account of *any* class of terms? Second, is it the correct account of a particular given element of the mentalistic vocabulary, such as 'pain'? Third, is it a good account of *all* referring terms? To the first question, I expect that the answer is *yes*. At the very least, it seems important that some terms be able to refer to things in the environment independent of the (surely fallible) identifying criteria in terms of which they are first conceived. For example, if whales were initially supposed to be fish but turned out to be mammals, we ought not to suppose that pre-Aristotelian references to whales were simply vacuous. To the third question, I think the answer is certainly *no*. There are clearly terms for which the identifying criteria are *constitutive*, even if, upon investigation, they turn out not to really apply to the identifying instances. Consider, for example, geometric terms like 'parallel'. I might learn the term in conjunction with both an example of two actual lines on a wall and a definition in Euclidean terms. It might turn out that the two lines on the wall in fact would converge at some distance, yet my use of 'parallel' tracks the Euclidean definition and not the properties of the actual lines used as exemplars.

The crucial question, then, is whether terms in the mentalistic vocabulary work in the fashion described by NS. I think, in fact, that they do not. And indeed I think that Kripke and Putnam would join me in this assessment. This has, in fact, already been argued in chapter 2. There it was argued that NS is ambiguous between two rival interpretations:

Canonical Essentialism: Natural kind terms pick out essential properties of mind-independent natural kinds, and one could in principle arrive at a "canonical (re)formulation" of the sense of any concept whose constitutive inferences would parallel all of the determination relations of the property.

Opaque Essentialism: Natural kind terms pick out essential properties of mind-independent natural kinds, but (at least in some cases) their nature is to some extent *epistemically opaque*, so that there is not a canonical (re)formulation available wherein the constitutive inferences would parallel all of the determination relations of the property.

Canonical essentialism turned out to be a form of broad reductionism, and thus (in light of further problems with broad reductionism) to be untenable. Opaque Essentialism is consistent with nonreductive physicalism, but it is not at all clear that it is a good interpretation of the semantics of the mentalistic vocabulary. NS is arguably applicable only to the "fillers" of role-filler terms, but mentalistic terms like 'pain' are arguably either role terms or else not role-filler terms at all.

In short, NS is a viable way of rescuing nonreductive materialism by way of epistemically opaque identities *only* if mental terms pick out fillers of role-filler terms. But it seems rather evident that this is not how they function. And thus NS cannot be used to salvage nonreductive materialism.

6.7. The Mysterian Gambit and Cognitive Closure

Some nonreductive physicalists have seized upon the "Mysterian" views brought forward by Colin McGinn (1990) and a principle of "cognitive closure." In answering the question of how psychophysical relations might be necessary yet abidingly opaque, writers like McGinn and Thomas Nagel (1974, 1986) suggest an epistemological rather than a metaphysical answer. The epistemological opacity may be a consequence, not of anything about the *world*, but of some feature of *our minds* that renders us incapable of grasping some real and necessary determination relations: "A type of mind *M* is cognitively closed with respect to a property *P* (or theory *T*) if and only if the concept-forming procedures at *M*'s disposal cannot extend to a grasp of *P* (or an understanding of *T*)" (McGinn 1990, 3).

I shall, for purposes of discussion, distinguish between "Mysterianism" and "cognitive closure" in the following way. I shall use the word 'Mysterianism' for the view that some phenomenon P is not fully explainable in non-P terms. One can thus be a Mysterian-about-P without having any additional explanation of *why* P is not explainable in non-P terms. The "cognitive closure thesis" I take as a particular sort of account of Mysterian unexplainability, one that traces it to features of our cognitive architecture. The irreducibility of A to B is due to cognitive closure just in case (a) A is metaphysically supervenient upon B, (b) we are unable to understand why A is metaphysically supervenient upon B, and (c) our inability to understand its metaphysical supervenience is due to some feature of our cognitive architecture. These definitions are stipulative. I recognize that 'Mysterianism', at least, has a variety of uses in the philosophical literature. (Sometimes, for example, 'Mysterian' is used in a fashion that implies, not only that the mind is not comprehensible in terms of something nonmental, but additionally that it is not comprehensible in its own terms either.) My usage does not reflect all of these, but singles out one of them for attention.

The label 'Mysterian' is usually employed exclusively with respect to Mysterianism (in my sense) *about mental phenomena* such as consciousness. By itself, this amounts to little more than an endorsement of the claim that there

are abiding psychological gaps. Dualists can, in my sense, be Mysterians as well. Mysterianism can also, however, be combined with nonreductive physicalism. Indeed, I tend to see *all* nonreductive physicalists as being Mysterians to some extent, as they all acknowledge explanatory gaps in the form of failures of reducibility. I shall refer (again stipulatively, as we are again dealing with a word with multiple usages) to the combination of physicalism with Mysterianism as "emergentism." "Emergent" properties, on this definition, are contrasted with both "fundamental" and "resultant" properties. Fundamental properties are those that are not metaphysically determined by any other sort of property. Resultant properties are those that are determined by other properties in a fashion that is epistemically transparent. Emergent properties are those determined by other properties in a fashion that is abidingly epistemically opaque.

Both Mysterianism and the worries about cognitive closure strike me as eminently sensible. We have reason to think that there *are* abiding explanatory gaps. And once one thinks about the question of whether God or evolution is likely to design our minds so that they are capable of grasping the ultimate natures of everything in the world and the connections between them, one is likely to adopt a stance of epistemic humility on the question. It is, at the very least, a substantive empirical question whether minds like ours are built for that. Even Mild Rationalism—the thesis that the ultimate natures of things are intelligible to minds like ours (though not necessarily on the aprioristic grounds favored by the more robust Rationalists of the seventeenth century)—ought not to be taken for granted.

But in light of scientific pluralism, it strikes me as wrong-headed to be a Mysterian *only* about the mind, since there seem to be other abiding gaps as well. Indeed, it seems to me that the right position to adopt is a kind of "Mysterianism all the way down." Likewise, even if there are *special* problems in *self*-understanding, it seems dubious that our minds are built to completely understand all of the determination relations in the nonmental world either. (Or, again, it is at least a substantive empirical question whether our minds have such abilities.) Considerations of cognitive closure suggest a strategy for approaching the question of *why* there are widespread explanatory gaps between the sciences generally: namely, to see whether these might plausibly be understood as a consequence of features of our cognitive architecture as employed in the sciences. This suggestion is taken up in the next two chapters, and the metaphysical consequences of this "Cognitive Pluralism" turn is explored in chapter 9.

As for nonreductive physicalism, Mysterianism and considerations of cognitive closure help it in one way, but not in another. On the one hand, they provide a kind of consistency proof for how it might be that there could be metaphysically necessary determination relations that are abidingly opaque to us: namely, if this were a result of some fact about our minds rather than about the world. This also allows us to see the various explanatory gaps as being on a par: each of them (or at least a large class of them) might involve metaphysically necessary determination relations that minds like ours are unable to fully

grasp (whether due to a single limitation that is the root of all such cases, or different limitations for different cases). So I regard nonreductive physicalism as a *consistent* position, and one compatible with the evidence for scientific pluralism. On the other hand, a consistency proof is a frail reed on which to support a metaphysical claim such as physicalism. Mysterianism and cognitive closure would also provide rival views, such as dualism, with a way of understanding the nonmental gaps that is consistent with their position (though dualists would have to be "fundamentalists" rather than emergentists about *mental* properties).

Moreover, the more we adopt a stance of epistemic humility—the more we come to doubt that minds like ours are built so as to adequately reflect all of the empirical and necessary connections between things in the world—the more we will be inclined to be suspicious of things like our intuitions about metaphysical supervenience. The empirical evidence supports the view that at least many explanatory gaps involve *some* sort of strong relation between properties that cannot be reduced to one another. But physicalism, like dualism, goes beyond this, to favor a particular metaphysical interpretation, and indeed one positing the very strong modal relation of metaphysical necessity. Such transempirical claims cannot be adjudicated empirically. And if we adopt a perspective of epistemic humility, it is not clear that we should place too much weight on modal intuitions about such matters either, especially when these intuitions differ so markedly between well-trained thinkers. As a consequence, I regard the Mysterian gambit as providing a consistency proof for nonreductive materialism, but not as providing any reason for favoring it over dualism or a more radical pluralism.

6.8. Conclusion

Initially, nonreductive materialism seemed to be an attractive position in the wake of scientific pluralism. On closer examination, however, this initial impression is given the lie. The conventional strategies for filling out nonreductive materialism—Davidson's interpretationalism, token identity, causal closure, contingent identity, New Semantic identifications, Mysterianism, and cognitive closure—all fail to provide a grounding for the *physicalist* commitments of nonreductive materialism, or to recommend it above its competitors. The nonreductionism is vindicated by scientific pluralism. But the physicalism is left without an anchor.

PART III

Cognitive Pluralism, Explanation, and Metaphysics

This page intentionally left blank

7

Two Forms of Pluralism

A possibility little explored in recent philosophy of mind and metaphysics is that scientific pluralism gives us reason to embrace a more radical pluralism as a general philosophical stance, on a par with monism and dualism. Given that each of the familiar positions in philosophy of mind is considerably weakened by scientific pluralism, it behooves us to explore alternatives. Moreover, to the extent that we feel compelled to embrace the ontologies of the sciences, and these seem to involve several distinct ''regionalized'' ontologies that resist integration through intertheoretic reductions, there is a *prima facie* case for some type of ontological pluralism that needs to be examined.

Such a pluralism might take two very different forms, however. One of these is cast directly as a thesis about ontological *inventory*: that the world is composed of an irreducible plurality of kinds and properties. This view is represented most prominently (and almost exclusively) in contemporary philosophy of science by the ''promiscuous realism'' advocated by John Dupré (1993). Historically, it also includes Aristotelian metaphysics, which had separate substance-kinds for each biological species. The other form of pluralism is most easily understood as an extension of cognitivist/idealist or Pragmatist approaches to philosophy of science and metaphysics. Traditionally, cognitivists and Pragmatists have not been content to treat the ''inventory ontologies'' of the sciences, common sense, or (realist) philosophical metaphysics as metaphysical bedrock. Instead, they attempt to cash out notions like ''object'' in terms of facts about cognition (such as representational structures employed by the mind) or material and social practices (such as laboratory procedures and language). (I shall generally lump idealists under the ''cognitivist'' heading, in

that they explain things about epistemology and metaphysics by appeal to features of cognitive architecture. This is clearest in the case of the transcendental idealisms of Kant and Husserl, but can also be applied to Berkeley, for whom the status of "being" is cashed out in terms of "perception" or "perceivability.") The initial point of entry to this sort of philosophy is epistemological, though it has consequences for ontology as well, at least in the sense of "critical ontology": the study of the nature of being. Whereas realist pluralism sees scientific pluralism as pointing to a plurality at the level of ontological inventory, cognitive and pragmatic pluralism see it as pointing to something about the kinds of thinking and practice that are involved in our constituting a world of objects, both in ordinary thinking and in the sciences.

This chapter undertakes two tasks. The first is to explore the kind of "promiscuous realism" advocated by Dupré as an alternative to monism and dualism. The second is to develop a view I call *Cognitive Pluralism* as an explanation of the disunities found among the sciences. The next two chapters continue this discussion of Cognitive Pluralism, chapter 8 exploring it as a more general thesis about human cognition outside the sciences, and chapter 9 considering its implications for metaphysics.

7.1. Dupréved Pluralism

Why are there explanatory gaps between, and perhaps even within, the sciences? It is useful to pattern our answers after those given to another question: *Why are there explanatory gaps between mind and the world of nature?* To this question, reductionists answer that it is solely because of our current state of ignorance of the reductive relationships that are really there, only yet undiscovered. Dualists answer that it is because of an ontological gap between material and mental substances. And Mysterian advocates of cognitive closure answer that it is the result of what happens when minds like ours turn their attention to understanding themselves.

A similar set of answers offer themselves with respect to the question of why there are gaps between different natural-scientific domains. Reductionists maintain (a bit less plausibly now) that it is a result of our current ignorance of reductive relationships that are really there to be found if only we look in the right ways and in the right places. Dualists claim that the psychological gaps reflect a dualistic ontology, but have no distinctive answer with respect to the other gaps. Some nonreductive physicalists take the Mysterian approach with respect to the psychological gaps, arguing that these are due to problems encountered in understanding how the mind arises out of material processes, but again have no distinctive answer to offer with respect to the other gaps. A realist pluralism takes the dualist strategy and applies it more generally: the explanatory gaps reflect a prior and more basic plurality in ontology. Cognitive and Pragmatist pluralisms take the Mysterian strategy embraced by some nonreductive physicalists and apply it generally as well: explanatory gaps are a consequence of facts about how we represent and intervene in the world.

A realist ontological pluralism has been explored in philosophy of science, most notably in John Dupré's (1993) "promiscuous pluralism" or "promiscuous realism" and Philip Kitcher's (1984, 2003) "pluralistic realism." Both Dupré and Kitcher are concerned primarily with biological categorization, particularly notions of "species." This is a pressing problem in philosophy of biology, as there are a great number of such concepts at work in biology itself. (Biologist John Mayden [1997] distinguished twenty-two such notions!) Most prominently, some notions of "species" treat two organisms as being of the same species if they can interbreed to produce offspring that are themselves capable of breeding. On other notions, a species is a historical particular, in which lineage is crucial. To illustrate with an example (one that might offend hands-on philosophers of biology like Kitcher and Dupré), according to notions of "species" based on interbreedability, we and our twins on Twin-Earth would all be humans. But if species are historical particulars, then we and our twins are of different species because we lack a common ancestor. Both Kitcher and Dupré, of course, illustrate their point with careful expositions of examples from biology that do not depend on thought experiments.

The crux of the matter is this: different biological classifications cross-classify. Yet all of them, or at least a sizable number of them, lay good claim to being respectable scientific classifications. Kitcher and Dupré take a realist view of the situation, holding that each viable classification should be viewed as picking out legitimate kinds in nature. They suggest that we should not view alternative classification schemes as *rivals* for the honor of being the *right* scheme, but rather should take the crucial turn of rejecting the assumption that a plurality of classification schemes is any kind of barrier to taking the kinds picked out in any of them, or all of them together, as *real* kinds in nature. As Kitcher (2003, 128) puts it:

> However [realism] is developed, it will prove compatible with pluralism about species. *Pluralistic* realism rests on the idea that our objective interests may be diverse, that we may be objectively correct in pursuing biological inquiries which demand different forms of explanation, so that the patterning of nature generated in different areas of biology may cross-classify the constituents of nature.

On the question of *why* there is this diversity of classificatory schemes, both writers combine a Pragmatist theme (that the plurality of kinds is connected to the variety of the interests bound up in the sciences) with a realist theme (that they are rooted in real commonalities in the things studied by the sciences). Kitcher's pluralism is expressed solely within philosophy of biology—as a thesis about species, units of selection, functions, and genes—and is committed to realism only about categories that do work within the science. Dupré's pluralism is more radical, allowing culinary classifications and prescientific classifications within which whales count as fish to stand on a par with scientific divisions of species. This difference is significant in how their respective claims might be thought to have implications beyond biology. One could

accept Kitcher's pluralism with respect to biological concepts while denying that they had ontological implications beyond biology by *stressing* the realist theme that there are, in fact, many different processes going on in organisms and in species over the course of evolutionary history, and these require us to employ multiple cross-classifying schemas. But this may be a *peculiarity* of biology, one not shared by, say, physics or chemistry.[1] Dupré's pluralism seems to have much broader ramifications. Because Dupré is explicitly contrasting his promiscuous pluralism with reductionism and claiming that the ontological commitments, not only of the special sciences, but even commonsense practices, trump the imperialism of physics, his pluralism is truly "promiscuous," in that it cannot be confined within the boundaries of a single science, or indeed within scientific domains at all.

Both Kitcher and Dupré, however, are also antireductionists. (Indeed, Dupré takes on reductionism as one of his major targets, and as the primary foil for pluralism.) Thus even Kitcher's relatively well-contained biological pluralism must be taken up, not merely as a kind of canon for the scientific legitimacy of kinds, but also as a metaphysical position. (Dupré's metaphysical ambitions should be abundantly clear from the title of his [1993] book: *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*.) The case for *realism* here is a familiar one: namely, that the success of scientific explanations gives us reason to be realists about their theoretical posits. But if one combines this with antireductionism and insists on giving the resulting view a metaphysical interpretation, one is left with a choice between a Mysterian nonreductive physicalism and a nonphysicalist ontological pluralism that denies supervenience as well. Dupré (2001, 5) endorses the latter view:

> I do not deny that the physics of elementary particles may very well eventually provide us with the whole truth about something, namely the nature of the stuff of which the world is ultimately composed. I do not believe that there are, in addition to the things that physicists theorize about, immaterial minds or deities. I believe, rather, that there are countless other kinds of things: atoms, molecules, bacteria, elephants, people and their minds, and even populations of elephants, bridge clubs, trades unions, and cultures. I agree with the physicalists that to the extent that these things are composed of anything they are, ultimately, composed of the entities of which physicists speak. Where I differ is in my assessment of the consequences of this minimal compositional physicalism. The truth about physical stuff, in my view, is very far from being the truth about everything.

Kitcher (1984, 350) also seems to reject physicalism, or more specifically the causal closure of physics, and to view pluralism in biology as revealing a deep fact about the ontology of nature: "Antireductionism construes the current division of biology not simply as a temporary feature of our science stemming from our cognitive imperfections but as the reflection of levels of organization in nature." For Kitcher, the crucial issue in the transition from explanation to

metaphysics is that biological explanations in fact *explain* things that would be left unexplained by lower-level sciences. Commenting on the relation between molecular-level and cytological-level accounts, Kitcher writes: "The molecular derivation forfeits something important. . . . The molecular account objectively fails to explain because it cannot bring out that feature of the situation which is highlighted in the [biological] cytological story" (350). The implicit argument seems to be something like this: Case studies in the sciences show that the special sciences explain things that are left unexplained by sciences studying their component parts. Thus the thesis of causal closure under physics is false. But if the thesis of causal closure is false, then physicalism is false, and the entities of the special sciences make causal contributions over and above those of simpler systems. This argument seems implicitly to invoke something much like Negative EMC (i.e., the principle that principled failures of reduction entail failures of metaphysical supervenience as well), as it argues from a failure of explanation to a metaphysical conclusion.

While Kitcher and Dupré arrive at ontological pluralism through case studies in biology, one could also motivate a realist pluralism in a more directly metaphysical fashion. If one combines a rejection of reductionism with Negative EMC, then one is presented with a forced choice between a radical ontological pluralism and a radical eliminativism. Either the kinds of the special sciences can be ontologically legitimate without supervening upon physical phenomena, or else we must accept an eliminativist stance toward them. To the extent that one accepts the customary realist premise that successful scientific practice gives us reason to be realists about the theoretical posits of the sciences, irreducibility thus leads one to a pluralist ontology.

Realist pluralism, however, has some obvious drawbacks. Foremost among these is that it seems highly counterintuitive. Even Dupré seems willing to embrace a "minimal compositional physicalism" that holds that "atoms, molecules, bacteria, elephants, people and their minds, and even populations of elephants, bridge clubs, trades unions, and cultures" are "ultimately, composed of the entities of which physicists speak," at least "to the extent that these things are composed of anything." But most of us are inclined to assume that if, say, molecules or elephants are *composed* of nothing but whatever the basic physical particles are, their behavior must be completely *determined* by the properties of those basic physical parts. And if we have reason to believe that the science of the basic physical parts cannot explain everything about molecules or elephants, we are inclined to seek a way to hold on to the ideas of supervenience and causal closure and interpret the failure of understanding in some other fashion, say, in terms of Mysterian-style *cognitive* closure. That is, when awakened from our reductionist slumber, we have a tendency to try to retreat strategically to the view that, though physical *theory* may not *explain* everything explained by the special sciences (an epistemic gap), the physical *laws themselves* must *entail* everything covered by the special sciences, even if we cannot understand why or how they do so.

Here we have the Mysterian move in a general form, not confined to psychology, a development that I laid the groundwork for in chapter 2's

taxonomy of positions in philosophy of mind. In this view, radical pluralism has a rival that I think Dupré and Kitcher have failed to anticipate: a physicalism combined with "Mysterianism all the way down." This is not, indeed, the view physicalists have traditionally been wont to take. But to the extent that one is attracted by nonreductive physicalism in philosophy of mind, a physicalism that is nonreductive with respect to the objects postulated by other special sciences should not be so great a stretch. It is simply the same move—asserting metaphysical supervenience while denying epistemic transparency—writ large. If it is good for the psychological gander, it should be good for the chemical goose, and all the biological goslings to boot.

In saying this, I do not mean to *commend* nonreductive physicalism, as should be evident from chapter 6. My point, rather, is that when we take observations about irreducibility from the philosophy of science and then move them to the realm of metaphysics, we experience a clash of intuitions. One intuition is that explanatory gaps, if they are principled and abiding, entail ontological gaps as well. If you can't *explain* B in terms of A, there must be something going on with B beyond what is present in A. (This is Negative EMC.) But this is offset by the countervailing intuitions (a) that molecules, elephants, and the rest of the objects of the special sciences (with the possible exception of the mind) are composed of nothing but the objects of basic physics, and (b) that the facts about the whole cannot involve anything over and above what is determined by the facts about the parts. We might call this the "intuition of determination by composition" or "compositionalism."

Dupré at least *seems* to embrace the first compositionalist intuition: that molecules and elephants are composed of nothing but whatever the basic physical parts might be (though his hedge "to the extent that these things are composed of anything" may signal otherwise). He thus seems to embrace an emergent realism, in which composite kinds can be "independent" by dint of additional *causal* powers, denying the second compositionalist intuition. The kind of thoroughgoing Mysterian nonreductive physicalism I have outlined, by contrast, denies Negative EMC. The price of this is not negligible, as it amounts to giving up on the assumption that the physical universe is fully intelligible.

Here we are brought to an important crossroads. We have two sets of intuitions—Negative EMC and the two compositionalist intuitions—that enjoy some measure of initial plausibility. Of course, not everyone feels these intuitions as having equal pull. Indeed, some may lack one or more of the intuitions altogether. Thus physicalists and pluralists may look at the same body of evidence and each feel that there is a clear implication for ontology, yet disagree completely on *what* implication it might be. My sense is that, if one were to count heads, one would find that the physicalist and compositionalist intuitions have considerably more advocates than their pluralist counterparts. But this is merely a sociological fact. It may reflect only passing intellectual fashion, or perhaps the comparative strengths of intuitions that may be grounded in some fact of our cognitive architecture, yet be illusory. Moreover, it is not clear that science itself can provide any guidance as to what road to take. The pulls in either direction come, not from an assessment of what is

present in the science—both sides can agree on that—but from competing *intuitions* that go beyond the science.

How are we to choose which road to take? Indeed, is there some alternative path we may have neglected? To progress beyond this point, we need some standpoint that gives us a way to assess whether any of the intuitions that pull at us are trustworthy. And this requires us to look at our scientific practices, and our transempirical intuitions, from a cognitive standpoint. As it happens, doing so will also provide an alternative route that has much to recommend it.

7.2. Interlude: Four Agenda-Setting Questions

Over the course of the past several chapters, we have accumulated a number of new issues that seem to loom large on the post-reductionist landscape. It is worth making these issues explicit, and as framing agenda for post-reductionist philosophy of mind:

1. What is the origin of scientific plurality and failures of reductive explanation?
2. Is the Negative EMC tenable, and indeed how would we go about deciding such a question?
3. Are intuitions to the effect that the natural phenomena encountered in biology and chemistry supervene upon basic physics motivated apart from discredited reductionist arguments?
4. Is the intuition that there is something different and special about the psychological gaps defensible, and if not, can it be explained away?

None of the positions thus far surveyed—reductionism, eliminativism, dualism, nonreductive materialism, or promiscuous pluralism—has provided the resources to address these questions. Chapter 5 developed some strategies for addressing the fourth question on behalf of the dualist, and the first part of this chapter supplied the promiscuous pluralist's attempt to answer the first. In the remainder of this book, I shall recommend an alternative view that has some claim to addressing all four.

7.3. Cognitive Pluralism and Philosophy of Science

The position I wish to outline here I call Cognitive Pluralism. It is "cognitivist" in that it traces features of our *understanding* of the world to features of our *cognitive architecture*, that is, to empirical facts about how minds like ours model features of the world. It is also, in some ways, a *Pragmatist* position, in that the nature of our explanatory interests and of our interactions with the world also plays a role in determining the form of models in science and outside it.[2] It is a *pluralist* position in that it holds that we relate to the world through an irreducible plurality of special-purpose models that are not

reducible to a single common denominator or unifiable into a single axiomatic system. Indeed, if a worldview is anything like an all-encompassing axiomatic system, we do not have anything so global as a worldview at all. Instead, we triangulate the world by deploying various models, each of which is good enough for particular things. In some cases, like successful scientific theories, these models are very good indeed, and apply very broadly. However, the partial and idealized character of these models, *qua* cognitive models, poses barriers to their integration into a single supermodel. In short, reductionism is not just a thesis about how the *world* is, it is a thesis about what the *mind is like* as well: namely, one that assumes that our various special-purpose representations can be regimented into the form of a single axiomatic system without loss of content or explanatory power. This hypothesis has two unfortunate characteristics: it is highly unlikely that God or evolution *would* build a mind like this, and it seems to be empirically false.

I shall begin by setting out Cognitive Pluralism as a thesis in philosophy of science, designed to explain the disunity of the sciences. It is not, however, a thesis *only* about the sciences. Rather, it claims that scientific thought is but an especially exact and regimented case of an activity of *cognitive modeling* that is employed much more broadly. (Some elements of it are very likely even shared with other species.) This claim will be argued in chapter 8, where I argue additionally that this thesis about cognitive architecture is plausible on a number of independent grounds, and indeed draws together seemingly disparate insights about mental modularity, domain-specific reasoning, and scientific disunity.

As a *cognitivist* position, Cognitive Pluralism bears some similarities to the transcendental idealisms of Kant and Husserl, and at the same time shares continuities with the strand of naturalistic epistemology that avails itself of ideas from cognitive and biological psychology to illuminate questions about knowledge and understanding. How much Kant would have liked the degree of disunity I am advocating is a matter allowing of some dispute.[3] However, I also follow Kant (in chapter 9) in taking seriously the metaphysical implications of this approach, claiming that it sheds light on questions of critical metaphysics, such as what it is to be a thing. (It is, to the best of my understanding, largely neutral on questions of "positive" or "inventory" ontology, with the possible exception of supposing that there are such things as minds.)

The main points I wish to argue in the following sections can be summarized as follows:

1. Scientific laws and theories are models of particular aspects of the world.
2. Such models are products of cognitive processes of model*ing*, and hence their form is determined in part by human cognitive architecture.
3. Scientific models are *idealized*, and indeed can involve several forms of idealization.
4. Each model must employ some particular representational system in describing its subject matter.

5. Both the idealization of models and the choice of representational systems can present barriers to the integration of different models and may necessitate the partial rather than comprehensive character of individual models.
6. Empirical facts about human cognitive architecture will constrain the types of models we can conceive, understand, and employ, and there can be disunities within the sciences that are artifacts of our cognitive architecture.

In the next chapter, I argue that these features are not confined to *scientific* models, but are features of mental models in general. Indeed, the features of scientific modeling, I argue, are merely a special case of what seems to be a general principle of human cognitive architecture.

7.4. Scientific Disunity: A Cognitivist Perspective

Scientific theories are models of particular aspects of the world. Events in the real world are complicated, messy, and difficult to understand with precision. One of the great achievements of modernity was to hit on a method of isolating different natural invariants like magnetic and gravitational force, both intellectually and through careful experimental arrangements, so that they can be studied more exactly in isolation, or at least in a situation as closely approaching isolation as our instruments will allow. Modeling thus involves *abstraction*, or bracketing off some features of the world in order to attend to others. Abstraction is a bit like looking at something through a special lens that filters out some features so that others stand out more clearly.[4] Sometimes when one does this, what one "sees" are really artifacts of the experimental apparatus or representational system. But sometimes one hits on what seem to be real invariants that are not artifacts of our experimental setup, and this can lead to laws and theories. The resulting laws and theories capture invariants that were not apparent before and are at least sometimes cast in a form that is mathematically exact.

But this insight and exactitude is bought at a price: one is observing reality through a cognitive filter, and as with all filters, some things are filtered *out*. As a result, reality-as-modeled does not behave exactly like objects in vivo (or at least it does not do so except under very special conditions, such as what Cartwright [1999] calls "nomological machines"). Things do *not* really fall according to Galileo's laws, because other forces like friction, wind resistance, and electromagnetism, which are filtered out of the gravitational model, are at work in vivo. Additionally, many filters *distort* their subject matter in representing it. A red filter on a telescope or camera may bring out things in a scene one did not see before, but it also distorts the image in the process. Analogously, the cognitive filter involved in a particular scientific model might distort features of the world, say, by treating bodies as point masses or collisions as perfectly elastic.[5]

Abstraction is not a bad thing. Indeed, a major advance of early modern mechanics was that it abandoned the Aristotelian approach of explaining

motion in terms of the specific nature of the moving thing. This allowed the insight that *bodies per se* (a notion unavailable to Scholastic science) fall in a particular way, regardless of whether they be cabbages or kings. Perhaps more problematically, it also seemed to license the postulation of a generic type of substance called "matter." But some abstractions are significant in ways others are not. Gravitational models, for example, abstract away from both color and shape. But abstracting away from shape has ramifications that matter in a way that abstracting away from color does not. The *color* of a ball dropped from the Tower of Pisa is utterly irrelevant to its trajectory in freefall. But its *shape* is not irrelevant: a piece of paper wadded into a ball falls very differently from the same piece of paper folded into a paper airplane, and a live bird dropped from the tower behaves differently from a dead one. I use the term '*idealization*' to refer to abstractions that matter for predictions of real-world behavior. (And of course what matters may depend on just what you are trying to predict. The head of Charles I no doubt rolled off the block much as a similarly sized cabbage would have, but the latter would presumably not have caused a bloody civil war. Abstracting from regalness is innocent in ballistics, but not in politics.)

7.5. Three Types of Idealization

Idealization is an important and generic feature of scientific models, including those of physics. As astrophysicist Paul Davies (2004, 11) writes:

> Physics is predicated on the assumption that the fundamental laws of the universe are mathematical in nature. Therefore the description, or prediction, of the behavior of a physical system is implemented by mathematical operations. These operations are necessarily idealizations; for example, the use of differential equations assumes the continuity of space-time on arbitrarily small scales, the frequent appearance of numbers like π implies that their numerical values may be computed to arbitrary precision by an infinite sequence of operations.

Idealization of scientific models is not, of course, *confined* to models of physics, or of other highly mathematized sciences. It is a feature equally important in the models employed by sciences like biology, in which mathematical laws play a much smaller role than they do in physics. Moreover (as the two examples in the Davies quote suggest), there are a number of different *types* of idealization at work in different scientific models, and these each introduce their own sorts of problems for (a) the relation between the model (or reality-as-seen-through-the-model) and the seething complexity of the real world and (b) the prospects for unifying a disparate set of models with one another.

There are probably quite a number of types of idealization that may prove significant for philosophy of science. Here, I shall discuss three types, which I call "bracketing," "distorting," and "approximating" idealizations, respectively.

7.5.1. Bracketing Idealizations

One type of idealization, arguably involved in *any* scientific model, is the *bracketing off* of other phenomena that may be at work in vivo. A gravitational theory, as such, idealizes away from electromagnetism, and vice versa. In a way, this is a fairly bland observation: that models are generally models of *something* but not of *everything*. But though this fact may at first appear bland and uninteresting, if we look at the matter historically, we will see that arriving at models of particular features of the natural world through bracketing idealizations was arguably one of the great achievements in early modern science. Both Platonists and Aristotelians had despaired of the possibility of a mathematically rigorous study of nature: Platonists because they viewed matter as fundamentally irrational and resistive to perfect instantiation of the Forms, and Aristotelians because they assumed that a science of motion or change would have to depend essentially on the specific nature of each individual type of thing—for example, each species of plant and animal—and hence there could be no generic and exact understanding of the natural world, but only a piecemeal collage. It was thus a great achievement when early moderns like Galileo and Descartes suggested the possibility of isolating mathematically describable physical principles that would apply to all physical objects, or at least to large classes of them, at one sweep. This moved the study of nature beyond the Aristotelian assumption of an irreducible (and large) plurality of types of "motion" and opened the doors to many subsequent unifications of scientific explanations.

It also *partially* addressed the Platonists' concern about the unruliness of matter. Platonists had identified matter itself as the source of disorder and irrationality. Bracketing idealizations allowed scientists to uncover strands of mathematical order *within* the material world. Platonists were *correct* in their concern that individual mathematical models fail to capture the complexity of real-world events. But they were too quick to assume that the reason for this was that matter introduces an incoercible "surd" into the mixture that renders it insusceptible to mathematical understanding. It is possible, through idealization, to gain rigorous understanding of principles that are both *real* and *physical*, such as gravitation, even though the resulting models underdetermine the complexity of real-world kinematics. The gravitation law captures something real and deep, and captures it with mathematical exactitude, even though it does not license descriptions and predictions of real-world behavior with comparable exactitude.

Bracketing idealization thus buys us insight into deep invariants in nature at the cost of driving a wedge between this *theoretical insight*, on the one hand, and exact description and prediction, on the other. A purely gravitational model of a body's fall adequately captures the contribution of gravitational force to the body's actual fall, and captures it in mathematically exact form. Were there real-world situations, or possible worlds, where gravity was the only causal factor in play, a gravitational model would be sufficient for describing and predicting how bodies would actually fall in such conditions (modulo concerns about

computational tractability, to be discussed below). But in the real world, there are few if any cases in which only gravity is in play, and hence the very bracketing move that allows us to isolate gravitational force for purposes of theoretical insight also results in a disparity between the behavior-of-the-falling-body-as-seen-through-the-model and how it will actually fall in the real world, where other factors like aerodynamics and electromagnetism are also at work.

Sometimes philosophers have been tempted to interpret this bracketing idealization of physical models in terms of the *domain* to which the model is applicable: either (a) that the gravitation law is a *ceteris paribus* law, applying exclusively to cases where *only* gravitational factors are in play, or (b) that it is a law that is *true* at "ideal worlds" where only gravitational factors are in play, and only "approximately true" (which is to say, strictly false) at the real world.[6] These interpretations, however, do not do justice to the role that idealized laws play in the sciences. It is not clear, for example, that there are *ever* any cases in the real world in which gravitation is the only force in play. "Other things" are *never* "equal" in the fashion needed to usefully interpret the gravitation law as a (true and nonvacuous) *ceteris paribus* law. If it is interpreted this way, the law has no true substitution instances. Nor will we do justice to scientific practice if we treat the gravitation law as making claims only about "ideal worlds." Scientists take such laws as saying something, and indeed something *true*, about the *real* world. Of course, they are never so naïve as to assume that *what* the law tells us (and tells us truly) is how real-world objects will *always actually behave*, in complete exactitude. The law is an *idealized* claim about the world, that is, about the world as seen through an idealized model that isolates one deep physical invariant.

Given that we understand the world scientifically through individual idealized models of different features and cases, it is natural to ask how these models can then be recombined. There is no *single* answer to this question. The answer in a given case depends crucially on other features of the models in question.[7]

In a best-case scenario for reintegration of models, we would have a number of idealized laws that were *fundamental, independent,* and *jointly exhaustive,* and that could, as a result, be factored and recombined in such a fashion that the complexity of real-world interactions could be recaptured through vector algebra in the summation of forces. Models of gravitation and strong/weak/electromagnetic forces are generally regarded as meeting the first two criteria of being fundamental and independent of one another.[8] Our commitment to the individual aptness of these models leaves open the question of whether they are jointly *exhaustive* of the causal forces at work in nature. And yet, even with these two forces, and even when combining multiple component forces of a single type, there are problems in combining them through vector algebra.

One type of problem, recognized by Newton, was that the equations for combining forces, even when (thought to be) fully deterministic, can yet prove computationally intractable. Indeed, such a problem arises in the summation of gravitational interactions alone when the number of bodies involved is

greater than two. This is a simple case of classical chaos, where any finite approximation of a system will be too coarse-grained to prevent the generation of significant errors in the prediction of the evolving kinematics of the system. To calculate the motions of bodies influenced by gravity, one must break the time continuum over which gravitational influence is exerted into finite units, and any such finite approximation of the evolution of the system through artificially quantized time will diverge from a continuous evolution. The causal contribution of classical gravitation is understood to be *deterministic*, but there is no general method to turn the gravitation law into a perfectly exact prediction of real-world interactions over time.

7.5.1.1. Exorcizing Laplace's Demon. Laplace famously attempted to lessen the impact of this gap between the (assumed) pristine exactitude of the idealized model and its inability to yield comparably exacting predictions of real-world behavior through his thought experiment involving a Demon whose mind was freed from our computational limitations. Given the suppositions of the thought experiment, Laplace's Demon could know the position and velocity of every bit of matter in the universe, and could make calculations to an arbitrary degree of accuracy, thus leaving it in a far better epistemic position than any human trying to perform calculations on textbook problems involving two or three bodies. It would also know all of the laws of nature, thus freeing it from the question of whether there are still undiscovered principles, which is constantly in the background for human scientists. Laplace suggested that such a Demon could exactly predict, on the basis of its knowledge of laws and initial conditions, the subsequent state of the universe at each succeeding moment for the entire history of the universe. In contemporary terms, one might say that Laplace attempted to reposition the gap between theoretical exactitude and predictive approximation as a failure of *performance* rather than of *competence*. That is, the limitations of human scientists are traceable to a combination of (a) their limited knowledge of initial conditions and of the laws in play, and (b) their finite capacities for calculation. But one could remedy these problems through an unbounded application of the kinds of knowledge and abilities humans really possess, without needing to resort to some very different kind of mental ability, such as the "intellectual intuition of the noumena" that Kant credits to God but denies to human beings.

But Laplace was mistaken in his assessment of the situation. Or, alternatively, his account is ambiguous between two interpretations of how the Demon is supposed to predict the evolution of the system. On the natural interpretation, the Demon does so in much the way human scientists (supposedly) do when utilizing a technique like Euler's method: by assessing the forces acting upon each body at a moment of time and then predicting each body's resultant motion through vector algebra. Such a process would then be iterated for each succeeding moment of time. The problem is that *the notion that there is a "next moment" in time is very likely a philosophical fiction*. If the time over which gravitational interactions operate is continuous rather than quantized, there are no smallest units into which to divide the continuum.[9]

And in chaotic systems, any divisions of finite size are enough to generate error. No increase in computational performance can remedy this problem completely, though no doubt the Demon's predictions would be much better than our own. The limitation stems, not from our limited ability to apply techniques of calculation, but in the relation between those techniques and the theoretical law to which they are related.

The alternative interpretation of Laplace's Demon would be that it might have techniques very different from ours for making predictions based on the same set of laws and initial conditions. It is not clear exactly what this would amount to, as by definition it involves postulating that the Demon has some sort of mathematical techniques available to it that are unavailable to us. We might allow that *if such a Demon had techniques for turning laws into predictions without resort to finite approximations*,[10] then it might be able to predict the causal history of the universe with complete precision. But this leaves open the question of whether such techniques are really possible, even for a mind freed of some of the constraints of human minds. Perhaps more important, if such techniques were available to a more powerful mind, they would still be very different from what goes on in what we call "science," and it is not clear that we should draw any conclusions for real science from our speculations on what such an imaginary Demon might be able to do. If the thought experiment is *not* simply a way of talking about the ideal extension of what we do in science, it seems a perilous speculation, based on extrascientific intuitions. If it *is* simply a way of talking about the limiting case of scientific understanding like our own, what it leads to is a scenario yielding predictions that are inexact but asymptotically close to real-world kinematics, and that only in nonchaotic cases. (And since classical chaos is now thought to be a far more prevalent feature of the physical world than was imagined in Laplace's day, it is not clear that this takes us very far at all.)

7.5.1.2. Problems in Combining Idealized Models. We are also often at a distance from our best-case scenario when we try to *combine* different idealized models. Our models of gravitation and strong/weak/electromagnetic forces may generally be treated as fundamental and independent, but it is not always possible to combine their results in a coherent way. In particular, taken jointly, they generate inconsistent or incoherent results when applied to very dense matter. This is a problem that vexes theoretical physicists greatly. Indeed, it is the main problem driving the search for a unified field theory today.

When we turn to models of features of the world that are *not* independent of one another, things get much more complicated. According to current scientific models, gravity does not affect electromagnetism, strong or weak force, or vice versa. Their contributions to system dynamics are independent. But it is very different with models employed in other domains, such as biology and neuroscience. There, one is generally modeling systems in which complicated feedback cycles play a crucial role. (Indeed, such feedback loops become important in the physics of solid-state systems and condensed matter as well.)

Consider the types of examples one finds in neuroscience involving the formulation of models of the dynamics of neural areas that contribute to identifiable cognitive processes like figure-ground separation. Such models will generally attempt to provide mathematical descriptions of the causal relations among a small number of areas of the brain, say, layers of cells in the LGN, V1, and V2. Such models characteristically *ignore* (i.e., bracket) a much greater number of connections known to exist between these areas and other areas of the brain. Often these connections, such as thalamocortical loops, are known to play an important role in the real-world performance of the task being studied. Sever or inhibit these connections, and the subject can no longer perform the task.

This situation is in some ways analogous to the bracketing of other forces in the formulation of a gravitational model: the scientist is bracketing some causal factors to bring out the regularities of others. But in another way the cases are importantly disanalogous: whereas electromagnetism is causally irrelevant to the operation of gravitation, the thalamocortical connections that are bracketed in a neural model are, very likely, crucially important to the proper functioning of the areas being modeled. Such bracketing may still be *necessary* in order to isolate a neural "circuit" for purposes of modeling, but it carries with it a much higher price tag, in the form of an additional gap between the model and real-world behavior.

One might put the matter like this: if one were to create a laboratory situation in which all of the features bracketed by gravitational models were truly absent, objects in that environment would behave exactly as the gravitational model would entail. But if one were to create a laboratory situation in which the layers of brain tissue being modeled were causally isolated from other parts of the brain, those layers of brain tissue would not perform their functions as the model entails. Rather, their behavior would more closely approximate that of monkey brain sushi or Dr. Lechter's box lunch in the last scene of *Hannibal*. In short, the technique of factoring and recombination does not work so well in cases in which the system modeled is not truly independent of the forces that are bracketed. (Of course, a feedback system also involves problems of the computational sort already discussed. Indeed, feedback loops *among* elements present in the model can cause nonlinear behavior to pile up quickly.)

Additionally, the technique of factoring and recombination through vector algebra is available only in the case of *quantified* models. Many of the special sciences, such as the life sciences, employ many models that are not based in quantified laws, but in structural models or process models. Scientists can often move between models in a fashion that affords an informal integration of the information contained in each, but in a fashion that lacks the mathematical exactitude found in physical laws.

7.5.2. Distorting Idealizations

There are additional types of idealizations that are involved in many scientific models. One class of these consists in what I call "*distorting* idealizations." These occur when, instead of simply ignoring a property of a system we are

describing by bracketing it, we actually model it, but in a fashion that misrepresents features we know to be present in the real-world case. For example, in gravitational models, objects are often treated as point masses. The spatial properties of objects are not simply *absent* from the model—the model *does* model spatial relations; instead, the sizes and shapes of bodies are intentionally misrepresented. Similarly, in contact mechanics, collisions between particles are often treated as being ideally elastic. In economics, people are treated as ideally rational decision-theoretic agents. Whereas bracketing idealizations bracketed off separate forces, leaving them *outside* the model, these distorting idealizations simplify or massage phenomena that are treated *within* the model in a way that results in their being characterized in a fashion that is known to be false. To treat objects as point masses, collisions as ideally elastic, or people as ideally rational is to treat them as though they have properties that we are fully aware that real objects, collisions, and people do not have.

7.5.2.1. THE MATHEMATICS OF DISTORTING MODELS AND PROBLEMS FOR INTEGRATION. As Mark Wilson (2006) points out in his fine recent book *Wandering Significance*, the *mathematical apparatus* often plays a crucial role in the choice of models. No one would choose to use a model requiring cumbersome partial differential equations for a problem in contact mechanics or fluid dynamics for cases in which a model using linear or simple differential equations is available and adequate to providing solutions within an acceptable margin of error. Indeed, not only the *choice* of models but also their *development* is often driven by the mathematical machinery that is available and adequate to the problem. But as Wilson's extended examples repeatedly illustrate, textbook presentations of such problems tend to create the false impression that all cases of, say, billiard ball collisions or the flow of a fluid can be handled by a single model. Worse still, the different models employed sometimes turn out to be inconsistent with one another in ways that philosophers are likely to find disturbing.

Consider the case of billiard ball interactions. Generally, these are presented in textbooks in terms derived from Newton, sometimes supplemented by contributions from Euler. In these models, billiard balls are treated as rigid bodies, whereas real billiard balls distort under impact. Moreover, "simple equation counting readily establishes [that] the technique does not provide enough data to resolve what happens in a triple collision" (Wilson 2006, 180). As a result, the illusion is fostered that Newton's equations provide an adequate model for collisions generally, when in fact they provide only "accounts that work approximately well in a limited range of cases, coupled with a footnote of the 'for more details, see . . .' type" (180).

Now one might expect that what this signals is merely that *simple approximations* are employed for the cases where they provide sufficiently accurate results, but that when one turns to more complicated cases, one is forced to move to a more fundamental and adequate model, one that either (a) adds details that were left out of the simpler models, or (b) expresses a more general and fundamental principle, of which the simpler model is a special case that

can be handled by less taxing methods. However, as one follows the trail of footnotes, this is not always what one finds. In the initial treatment, one follows Newton's strategy of dividing the collision into two stages: one before and one after the collision, *without* covering the instant of collision except with an empirically useful kludge involving a "coefficient of restitution":

> Derived from Newton, the basic trick is to almost—but not completely—cover the history of our colliding balls with two descriptive patches, one devoted to the balls as they approach the collision and the other as they scatter away from it. But the actual events of compression and reexpansion that occur when our two balls contact one another are set within a little window that our method does not attempt to describe. Instead, we bridge over this temporal hiatus by matching our incoming and outgoing sheets according to a rule of thumb involving gross energetic qualities and a crudely empirical *coefficient of restitution* (in the simplest—and most inaccurate—treatments, one simply assumes that the balls are "elastic"). The rough reasonableness of such approximation can be justified by Riemann-Hugoniot style considerations, but it is plain that our method collapses the central causal events into an untreated temporal singularity. Notice how all the moments in which real spheroids display distortion have been swept into the collision singularity: Newton's treatment doesn't provide a whisper of a suggestion that billiard balls might be flexible. (Wilson 2006, 190–91)

However, things quickly become more complicated.

> But in the long run, this approach is too crude to handle the blows encountered in, *e.g.*, sophisticated aircraft design, where an entirely new mathematical army (partial differential equations *et al.*) must march on the scene like cavalry reinforcements. As we saw, in many books, the first wave of this incursion follows a strategy devised by Hertz, that breaks histories of our colliding balls into discrete stages whose compressed states are assumed to relax into one another quasi-statically. But . . . this treatment merely represents a (very valuable) stopgap, for Hertz's recipe isn't adequate to substantive internal wave motion or truly violent impact, where shock waves form as well. (191)

The point here is not that these latter cases are deeply mysterious; there are mathematical techniques for addressing them as well. Rather, the point (or *one* point, at any rate) is that we have no single *general* model of collisions, but cover the various cases through a variety of "patches." Part of scientific knowledge then resides in understanding of the patches, but another part resides in knowing how and when to move between them as the explanatory or predictive situation demands.

What, though, about the assumptions, mentioned earlier, that the move to more complex models must either add something to the simpler ones or else

be a move to something more fundamental, of which the Newtonian model is a special case? It would be erroneous to say that the models employed to account for deformations are simply the Newtonian model, plus something extra, at least if that means something like a truth-functional combination. As Wilson points out, the properties of the collisions, as described by the different models, are not only *different*, they are *incompatible*: "Balls do not alter their shapes in the Newtonian accounts but they do in the other treatments; they do not transmit waves in the Hertzian picture, *etc.*" (2006, 191). To restate the issue in my terms, *If one interpreted such models as involving unidealized characterizations of their phenomena*, they would turn out to be describing incompatible states of affairs. The proper moral to draw, however, is not that they are rival models, but that they are idealized models, each of which elegantly draws out particular features of collisions under particular conditions. But the cost of this is that the idealizations employed result in models that are formally incompatible with one another, in the sense that they cannot simply be recombined by the use of logical connectives or vector algebra.

What about the other suggestion, that the Newtonian model is a special case derivable from more fundamental models? The Newtonian model may indeed be a *limiting case* of more complex models (e.g., where the values for deformation of the balls involved and the factors producing wave phenomena within the balls are set to zero). But a mathematical proof that B is a *limiting case* of A is very different from a *derivation* of a theorem B from an axiom A. If A is an axiom (in a system S), and B is derivable from A, then B must be true (in S) as well. But if B is a limiting case of A, there is no assurance that B is a fully accurate description of any actual phenomena falling within the domain of A. Indeed, as often as not, "limiting cases" are *ideal* limits that are never actually reached in any real *cases*. That is, when the limit can be approached but never reached, showing that B is an ideal limit of A entails that A and B are logically incompatible as (unidealized) descriptions of a real-world system.

7.5.2.2. MODELS, TRUTH, AND APTNESS. What I wish to say about such models is that each of them is "apt," a word I am using as a term of art, in particular cases. I prefer to say that models are "apt" rather than "true" for several reasons. First, I wish to be able to say that scientific *statements* about particular *events* are true or false, and the meaning (and hence truth value) of a statement in the sciences is determinate only once one has pinned down what model one is using. Models do not so much make claims about the world directly as provide a semantic framework within which statements can be made. (For example, Newtonian and relativistic mechanics define different geometries for space and time, and hence different possibility-spaces for *claims* that can be made about the location and motion of particular bodies.) Second, with respect to some models, such as the Newtonian model of collisions, no one would wish to say that the way it describes collisions (as not involving deformations, shock waves, etc.) is to be preferred in all circumstances to an alternative that accommodates these phenomena. But we still need

a success-term for such a model to reflect the fact that in some circumstances it is counted as a *good* model by scientists. Separating the "aptness" of models from the truth of claims (modulo the operative model) allows us to do this.

The aptness of a model consists in how suited it is to purposes of predicting, describing, and understanding a particular range of cases. This means, among other things, that the expression "Model M is apt" is grammatically incomplete. It is always a shorthand for "Model M is apt for pragmatic context C."[11] The "pragmatic contexts" relevant to scientific theories are various types of description, prediction, and explanation. In physics, where models tend to be highly mathematical, this may suggest that "aptness" is directly a relation between mathematical structures and real-world processes and events. But this would be misleading. The kind of fit between model and world that is of interest in scientific modeling is not simply some abstract Platonic formal correspondence, but one that is useful in the cognitive and pragmatic goals constitutive of science.

Concentrating on aptness rather than truth is also useful in examining how it is that models that are known to misrepresent features of the phenomena they treat of can still be viewed as apt models. The cases where application of the Newtonian approach to billiard ball collisions is deemed apt—that is, "good enough" for the purposes at hand—are still known to be cases in which deformation of the balls actually occurs. That is, the scientist or engineer who employs the Newtonian model knows full well that she is not dealing with a "special case" where the balls do not deform, and her reason for employing the model is not a misunderstanding of the nature of the properties of colliding billiard balls. Rather, she is content, and well-advised, to employ such a model *for purposes of prediction* when the departures from real-world behavior it introduces are sufficiently small to fall within an acceptable margin of error. But she is also content, and well-advised, to employ such a model *for purposes of understanding* for a very different reason. Unlike the treatment of the instant of contact by "coefficient of restitution," the abstraction away from flexibility is not simply a computational kludge. Even though the limiting case described by the Newtonian model may be one that never occurs in nature, the model, by dealing with the limiting case, brings to light features of collisions that would remain obscure if one insisted on attending to all of the complexity of nature.

Wilson (2006, 184) draws a similar conclusion in his assessment of *why* there is this multiplicity of models even for a single range of phenomena such as collisions:

> The macroscopic objects we attempt to treat in classical mechanics are enormously complicated in both structure and behavior. Any *practical* vocabulary must be strategically framed with these limitations firmly in view. To be able to discuss such assemblies with any specificity, our stock of descriptive variables must be radically reduced, from trillions of degrees of freedom down to two or three (or smoothed out to frame simpler continua).

7.5.2.3. IDEALIZATION AND COMPLEXITY: SCIENCE FOR HUMAN VERSUS ANGELIC MINDS. But one might well ask, in what ways is a model that has a few degrees of freedom "better than" one that has trillions of degrees of freedom? In certain respects, it would seem that for an ideal mind, like Laplace's Demon, a "master equation" that captured all of the complexity of nature in one fell (if long-winded) swoop would be preferable to a patchwork of less adequate, idealized models. Philosophers have often implicitly assumed that questions about the virtues of a scientific account should be freed from the empirical constraints of minds like ours, so that theories are "better" when they are *more adequate to all the phenomena*, even if this means they would be comprehensible only to God, Laplace's Demon, and beings of a comparable celestial order. But we are not asking questions about the kind of understanding that God or the angels might enjoy. We are asking *why real (human) science is disunified*. And here the limitations and peculiarities of human cognition are eminently relevant. For example:

- If it is a feature of human cognition that there is an upper bound (perhaps a fairly small one) to the number of degrees of freedom that can be present in a model if it is to be comprehensible to a human mind, then one would expect viable scientific models to be found among those that fall within this bound, regardless of the costs such models must pay in the form of bracketing and distorting idealizations.
- If simpler models are more comprehensible and/or more computationally tractable to human minds than more complicated ones, then one should expect human scientists to weigh trade-offs between empirical adequacy and generality, on the one hand, and simplicity and computational tractability, on the other. (Thus Occam's razor may turn out to be, not an aprioristic ontological principle, but an implicit heuristic for beings with our cognitive architecture.)
- If the human mind has better skills for moving back and forth between different models as the problem demands than at formulating more comprehensive models that treat more cases at the cost of theoretical and computational complexity, then one might reasonably expect a patchwork of laws instead of a single unified model.

Of course, noncognitive constraints are also relevant. On the one hand, there are practical constraints: How good a job does a given model do in fitting a given set of cases? This is in part a question of the relation of models to the world, and not simply a result of human cognitive architecture. But neither is it completely independent of our minds. After all, how we group cases according to practical interests, of which theoretical interests are an important if rarefied subset, is partly a function of *what our interests are*, and this is a fact about our minds. Indeed, one thing we have seen is that we seem to have *competing interests* for things like theoretical understanding, generality, simplicity, prediction, and control. It is not usually possible to optimize all of these

interests at once in science, and the set of interests with which we approach science drives the real trade-offs we make.

On the other hand, there may well be mathematical constraints: it is easy to *say* that an "ideal mind" like that of Laplace's Demon could understand the laws of nature in such a way as to license predictions of all subsequent states of the universe. But to do so is to engage in speculation about the kinds of mathematical and computational resources that would be available to such a mind, resources that far outstrip our own. It may well be that there *are no* mathematical and computational resources that would endow a mind with such an ability: the relation between mathematical models and the real world may *require* that such models be idealized, piecemeal, and many in number. And our "natural" assumption to the contrary may be an instance of what Kant called "illusions of reason"—that is, projections of ways we understand things to an unrealistic "ideal" limit. Moreover, it seems plausible to say that at least many of the idealized models employed in our sciences are *fruitful*, not only in providing techniques *useful* for prediction and control, but also in uncovering real invariants in nature itself that would otherwise lie buried in the complexity of events.

7.5.2.4. Metaphorical Transposition and Distorting Idealization. Thus far, I have dealt with distorting idealizations that are closely linked with mathematical models that afford simplicity and computational ease, and that can plausibly be seen as limiting cases of the mathematical models needed to handle more demanding cases. These represent what we might view as the most innocent cases of distorting idealizations. But there are other examples that are less innocent. Often, for example, models are coined in one domain by the metaphorical transposition of concepts and models from another domain. In cognitive science, for example, the mind has been modeled as a digital computer. To take an example from Wilson, Charles Navier patterned his model of fluids in terms of the Navier-Stokes equations on his previous work on viscous solids. Here the transpositions are not always innocuous, and the more so as they may go unnoticed.

In the case of the computer analogy for mind and brain, for example, the metaphor was first posed at a time when it was believed that neurons were digital (on/off) circuits, and the role of feedback processes in the brain, and their disanalogies from Turing- or von Neumann-style computation, were unknown or at least underappreciated. While it is still a contentious point whether digital computation can emerge from a neural network having the architecture of a human brain, at the very least the metaphor of the brain-as-computer can blind us to the architectural principles proper to the brain itself. The computer metaphor recommended itself on a number of grounds: (a) it was a model we *had* in hand, which was better than no model at all; (b) it was itself modeled on (a philosophical understanding of) certain forms of human thought; and (c) Turing showed that it was an admirably general and flexible framework, as any formalizable process could be implemented (or interpreted) computationally. Of course, the third factor could be applied to indefinitely

many nonmental phenomena in nature as well, as evidenced by writers who have suggested that we view the entire physical universe as a gigantic computer (Wolfram 2002). But even people who find this latter generalization of the computer metaphor far-fetched are often inclined to view computers as providing an acceptable essentialist picture of the mind, in part because of the second consideration.

If one is not careful, the mind-as-computer metaphor can lead us down the garden path to some fallacious conclusions. What Turing proved about digital computation was that any function that can be formalized can also be evaluated by a general-purpose computing machine. Combined with the notion that mental states are "functional states," this suggested that mental states (such as beliefs) and processes (such as reasoning) might also be formalizable, and hence (a) that things like reasoning might be implemented in a computer (the basic tenet of "Strong AI"), and (b) that it might be computational (i.e., syntactically based symbol-manipulating) processes that account for intentional states and reasoning in humans as well (the basic tenet of the computational theory of mind).

Here, however, we have a kind of "fallacy of idealization" (Horst 1996) that bears curious resemblances to mistaking a limit for a special case. In mathematics, the formalization of a domain (such as geometry) consists in an axiomatization with derivation rules sensitive only to the syntax of the formulas, requiring no understanding of the semantics in order to perform derivations and computations. This does *not* mean, however, that semantics is *reducible* to syntax: the formal properties of the system do not fully determine its possible interpretations (Horst 1996). Likewise, even if we allow that some mental states and processes have good functional *descriptions*, this does not mean that their nature as mental states and processes is exhausted by these functional descriptions. In Putnam's Twin-Earth examples, my brain states when thinking about H_2O are functionally equivalent to those of my twin thinking about XYZ. But, more important, there are ways of gerrymandering sets of objects and events so as to produce a "system" of events consisting of molecular interactions in a bucket of water that are functionally equivalent to both (Block 1978/1980). That is, the fact that a system S shares a functional description with a mental state or process is not itself sufficient to ensure that S has *mental* states or processes at all. (Compare: information theory and thermodynamics share a mathematical description, but entropic informational processes need not produce heat.)

Failure to observe these subtleties has led to some absurd claims, such as that thermostats have beliefs on the grounds that they are functional-state devices. At most, what functionalism and Turing's proof allow us to say is that mental states and processes share an abstract form with processes we can implement in computing machines. This is very different from the claim that mental states and processes can be derived from what is present in their abstract, idealized formal models. This, like our discussion of limit cases, is an example of how it is dangerous to mistakenly see a relation of derivability when in fact one is faced with something very different, such as a limit case or an abstraction to formal properties alone.

In the case of fluid dynamics explored by Wilson, Navier derived the Navier-Stokes equations for elastic *fluids* on the model of Navier's equations for an elastic *solid*, helping himself to Newton's F = ma equation, with 'force' in this case decomposed into factors including "viscous force" $\nu\Delta u$. However, in this transposition, crucial terms such as 'force' and 'particle' undergo changes of meaning that are not metaphysically innocent. In particular, in fluids, what is called a "particle"

> does not consist of just the same molecules at all times. The interchange of molecules between fluid particles is taken into account in the macroscopic equations by assigning to the fluid diffusive properties such as viscosity and thermal conductivity. . . . The same fluid particle may be identified at different times, once the continuum hypothesis is accepted, through the macroscopic formulation. This specifies (in principle) a trajectory for every particle and thus provides meaning to the statement that the fluid at one point in time is the same as that at another point in time. For example, for a fluid macroscopically at rest, it is obviously sensible to say that the same fluid particle is always in the same place—even though, because of Brownian motion, the same molecules will not always be in the same place. (Tritton 1976, 50, quoted in Wilson 2006, 158)

As a result,

> it was eventually realized (first by Maxwell, I believe) that some of this applied "force" upon our "particle" could not represent the application of any true force at all (e.g., attractions and repulsions exerted by neighboring regions), but instead must express net losses or gains of momentum occasioned when more rapidly moving molecules enter and leave the appreciable volume that our alleged "particle" actually represents. (Wilson 2006, 158)

Here we encounter a type of distorting idealization that presents an unexpected barrier to theoretical integration. To the extent that an integration of models of different phenomena, say, elastic solids and fluids, depends on the univocity of terms like 'force' and 'particle', subterranean changes of meaning in such terms as we move across models present a barrier to integration, as both reduction and truth-function combination require an equivocity of terms that has been lost in such cases.

7.5.3. Approximating Idealizations

I shall briefly canvass a third type of idealization, which I call *approximating idealizations*. One class of these has in fact already been discussed: the need for finite approximations of intervals of time when one uses theoretical models in the service of predictions or descriptions of real-world kinematics. There are, however, other types of approximating idealizations as well. One of these,

mentioned in the quote from Davies earlier in the chapter, is found in the use of mathematical constants such as π. π is an infinite decimal sequence, which means that any finite representation of π is, strictly speaking, inaccurate. This is arguably innocuous at the level of *theory*. In a theoretical formulation, the symbol 'π' can stand in for an infinite decimal sequence. But when one turns to computation, some particular finite approximation, such as "3.14" or "3.14159," must stand in for π. Sometimes, such a finite truncation of π will *matter*, in the sense that calculations based on that truncated approximation will produce results that fall outside an acceptable margin of error. Of course, what counts as an "acceptable margin of error" is a *pragmatic* consideration. But, importantly, we must distinguish between *models-as-theory* (in which constant letters can stand in for infinite decimal sequences) and "*practical models*," or *models-as-employed-in-prediction* (in which we must make use of a finite approximation). When we speak of "computer models" of real-world phenomena, for example, we are always speaking of "models" in the latter sense, ones that employ finite approximations of constants like π. In such practical models, the truncation of mathematical constants gives rise to an additional gap, and sometimes a significant one, between the model and the real-world phenomena that it models. Likewise, if two practical models employ approximations at different levels of accuracy, combining them will lower the degree of exactness of the resulting calculations to the lower degree of accuracy. This is an important limitation to computing the kinematics of chaotic systems.

There are also *physical* constants, such as the gravitational constant and the Hubble constant, that bring in similar concerns. If we treat gravitational acceleration for a body falling to the Earth as being equal to 32 ft/sec^2, we are employing a finite approximation of a constant with an infinite decimal sequence. (And, of course, the situation is only worse for the fact that gravitational force is not uniform over the face of the globe.) Again, such an approximation may not matter in some contexts. But what matters in a context is, again, a pragmatic consideration, and such a calculation is never a fully accurate reflection of the forces at work in a real-world situation.

7.6. Models, Representational Systems, and Connections

In addition to the idealized character of scientific models, I wish to draw attention to another feature models possess: each model must *represent* its problem domain in some particular way. Or, to put it only slightly differently, each model employs a particular proprietary representational system (and hence *not* any of a range of alternative representational systems).

7.6.1. Intertheoretic Dissonance between Mathematically Exact Models

In physics, the representational systems are largely characterized explicitly by the mathematical machinery employed. Classical and relativistic models of gravitation each represent points in space and time geometrically, but one

employs a Euclidean and the other a Lorentzian metric. We can, of course, make claims about the global aptness of either model for describing the universe. But more fundamentally, each model defines a system for posing more particular problems. Each provides, as it were, something like a grammar for describing an indefinite number of possible situations in mechanics. The mathematics of the system sets the general constraints within which such descriptions may be posed, and also provides laws constraining the temporal evolution of the system. The mathematics of the system thus constrains *what* can be conceived within the parameters of the model and *how* objects must be understood to behave as seen through the interpretive lens of the model.

Importantly, there is no way of having a scientific theory of space, time, and gravitation *without* employing some such mathematical model. The particular mathematics of a given model then determines *how* events in space-time are understood. And since the geometries for flat space and curved space employ incompatible axioms, the models are thus *inconsistent* with one another, in the sense that one could not combine their axiomatic bases without generating contradictions. This is a simple, familiar, and historically important example of a way that properties of the models we employ in science can themselves pose a barrier to certain types of intertheoretic integration. One model is not reducible to the other, in the Carnapian sense of being derivable from it as a theorem from an axiomatic base. Nor is there any more basic axiomatic base from which both can be derived. Indeed, if they were simply to be combined through conjunction, contradictions would result.

This does not, however, mean that there are not powerful mathematical relationships between classical and relativistic models of space-time and gravitation. Classical space-time is in some sense a "limiting case" of relativistic space-time. But this does not mean that there are possible worlds that are exactly described by both systems, or at least not worlds anything like the real world. What it means, rather, is (a) that regions of relativistic space-time approximate the flat geometry of classical mechanics at low speeds and low densities, and hence (b) that classical mechanics can often provide "good enough" approximations of relativistic mechanics in such situations. The geometry of a relativistic universe can never be completely flat, however, at least as Einstein conceived the matter. In an Einsteinian universe, one cannot have mass without curvature of space. One could only have a flat universe if there were no mass. And since Einstein sided with Leibniz over Newton in viewing space as defined by mass, rather than an independently existing plenum, no such space could, on his view, exist.[12] Regions of a relativistic universe can *approximate* a classical geometry, but (at least on Einstein's view) a truly flat universe is incompatible with relativistic mechanics.

This, however, does not prevent physicists and engineers from *using* a classical model to understand particular situations where the effects of relativistic phenomena are negligible, or to make good predictions in such cases by using classical techniques. Doing so, however, requires an understanding of the "idealization class" of the model: that is, knowing under what conditions it may aptly be applied. Classical mechanics is *not* an apt model for

accommodating all of the things we know about mechanics, nor is it apt for handling problems involving very large masses or very high velocities. These fall outside of its proper idealization class. But this does not mean that it is simply "false" in the way that we might rightly say that Ptolemaic cosmology or the Greek four-element theory are "false." Classical mechanics *does* "say true things" about gravitational invariants, even though what it says abstracts away from other aspects of gravitation that often matter in real-world situations.[13]

In the case of theories of gravity, we are presented with two models that bear a very special relation to one another, in that one (the classical model) is in some sense a "limiting case" of a more general model (the relativistic model). (This is so even if it could never be an *actual* case in a world with a relativistic geometry.) But we are in a different situation when we bring models of *different* phenomena into contact with one another. A pressing case in contemporary physics is the relation between general relativity and quantum mechanics, which unifies strong, weak, and electromagnetic forces. These are presently understood to be separate models of different and independent "fundamental" forces in nature, though theoretical physicists have aspirations to unite them by postulating more "fundamental" phenomena, such as superstrings. It is possible to combine insights of the two models in their application to particular physical problems through algebraic methods, as the forces are understood to be independent. However, in certain cases, involving very high density matter, combining the models generates results that make no sense. Importantly, this discovery is not the result of observations or experiments that reveal the universe working in bizarre ways. Rather, it is a result of exploring mathematically the implications of our two best models of physical phenomena in particular types of situations. The problems are generated by the mathematics of the models themselves. That is, the choice of these particular ways of representing physical phenomena, however well-supported by their individual adequacy to known phenomena, generates a kind of intertheoretical dissonance.

What are we to make of this situation? The fundamental point I would wish to stress is that it is a type of situation that can routinely arise when we employ multiple representational systems to describe aspects of a common reality. Each representational system is chosen for its aptness for a particular set of problems, but this does not prevent systems optimized for different problems from being in theoretical dissonance with one another. To the extent that human minds are constrained to understand the universe through particular mathematical models, rather than some direct and unmediated contact with "the things themselves," this possibility is always on the horizon. Hence such a *cognitive* constraint *may* consign us to theory pluralism and intertheoretical dissonance in any future science that human minds can attain to.

Such a principled scientific pluralism may not be *necessitated* by the fact that we approach different aspects of the world through different models. Whether this is so is in part an empirical question about what kinds of mathematical models we can devise for different physical invariants, how well these individually fit the range of phenomena to be explained, and

how the mathematics of the different possible models might interact with one another. It may be, for example, that relativity and quantum mechanics can be unified *mathematically* through a new formalism, such as those of superstring theories, even if these produce no testable predictions of their own and require postulation of entities whose evidential base consists entirely in the unifying power of the models that require them. However, the fact that our best-confirmed scientific accounts produce anomalies when we combine them ought at least to give us pause when we are inclined to assume, on aesthetic or philosophical grounds, that such a grand unification *must* be possible. This may, after all, be a Kantian dialectical illusion rooted in an *impulse* to unify, deeply embedded in the human mind, that such minds can never in fact *fulfill*. Even if there are real or possible minds that *can* attain to a grand unified theory of everything, there are also surely possible minds that have limitations that prevent them from doing so. The empirical question is *which class our minds fall into*. This is not just a question about the nature of the *world*; it is also, and perhaps in larger measure, a question about *our cognitive architecture*.

7.6.2. Intertheoretic Dissonance in Nonmathematical Models

It is not only physics that employs models with proprietary representational systems. However, the models employed in other sciences are often less completely mathematical, and in some cases are not quantitative models at all. Physical chemistry, for example, employs structural models of molecules. Darwin proposed a nonquantitative model of speciation through variation and selection (though successive generations added quantitative models ranging over populations). Chomsky proposed a model of acquisition of grammatical competence through hypothesis formation. And so on. Generally, however, such models have at least some elements of formal structure, even though this may not come in the form of the algebraic equations that play such a central role in physics. And even if they are initially proposed informally, such models can often be made more formally exact in nonalgebraic ways, such as through computer modeling of grammar acquisition in cognitive science, or the imposition of geometric and topological descriptions of molecules and the bonding between atoms.

Such cases, however, help to illustrate an aspect of scientific theory that is easily missed if one concentrates on mature areas of physics. It is possible to *picture* a curved two-dimensional space, and to build a physical model of it, without understanding it in terms of particular geometric axioms. But given the types of problems that mechanics is designed to address, it may not be possible to actually have a concrete *theory* of space-time that addresses such problems independent of a concrete mathematical model. By contrast, the kinds of physical models of molecules used in a science classroom *do* model salient properties of molecules, such as their rough geometric and topological structures, even if the ways those models are employed may underdetermine exact geometries or things like the flexibility of the bonds. Likewise, a general thesis that a cognitive process is ''computational'' underdetermines the precise

computational processes that could underwrite it. Yet in some ways it is the very ill-defined flexibility of such models that allows them to be used in furthering understanding. A loose and inexact model may help guide us in how to formulate the right questions to ask in order to end up with a more exacting model and to test our hypotheses. (This may often be true in the case of basic physics as well, though the steps that led to the mature theories we now possess may be lost in the retelling of textbook stories.)

Such cases also illustrate the role of *metaphorical transposition of concepts* from one model to another. Key theoretical ideas are often arrived at by taking concepts or even entire models from one domain and attempting to reconceptualize another domain through the same explanatory apparatus. Thus the atom might be conceived of on the model of the solar system, fluid interactions on the model of billiard ball collisions, or infraconscious processes like language acquisition on the model of conscious and explicit reasoning. Such metaphorical transposition of concepts and models is often fruitful, in the sense that it eventually leads to apt models of a new domain. But it also has perils that can easily be missed and that can lead to philosophical (and scientific) fallacies.

Take, for example, models of fluids based on the transposition of prior models of particle collisions among solids. One assumes that fluids are composed of tiny particles that are governed by laws of mechanical interaction using Newtonian equations. However, we saw earlier that, in the original context, "particles" were entities whose mass was unchanged over time and whose identity was constant. In fluids, however, the notion of a "particle" must be understood differently, as molecules can pass between fluid "particles" over time. The "particles" thus do not preserve their meriological identity over time, and the notion of "force" utilized has likewise strayed from its original moorings.

This kind of transposition of concepts between models, which Wilson (2006, 159) calls "property dragging," requires us to be vigilant when making inferences using words like 'particle' and 'force' that cross the boundaries between different models. If we fail to do so, we are at risk for paralogistic fallacies.[14] One approach to such situations, which guided many Positivist philosophers, is to view it as calling for a more exacting separation of terms (say, 'particle_1' and 'particle_2'). To leave it at this, however, is to ignore the project of understanding how notions like "force" and principles like $F = ma$ have actually crossed the boundaries between models fruitfully, in spite of such changes of meaning. (Wilson's *Wandering Significance* is an extended philosophical examination of the kinds of justifications of such transpositions that have helped underwrite them for the scientists, and I commend it to the reader who wishes to explore the topic in much greater depth.)

7.6.3. Open Questions

The employment of particular representational schemes can generate problems for theory integration, both in the form of mathematical inconsistencies and in the form of metaphysically significant changes in the meanings of

transposed concepts. But how deep and serious are these problems? Here it is useful to distinguish two layers of issues. The first is the issue of how two representational systems employed successfully in the sciences can be integrated—for example, whether they are formally inconsistent or generate paradoxes. The second is whether inconsistencies and paradoxes are *merely* artifacts of the representational systems employed or indicative of something deeper. To take a historical case, there was a dispute over the nature of light in classical optics, with experiments that seemed clearly to support the claim that light is a wave, and others that seemed just as clearly to support the claim that it is composed of particles. In *classical* optics, these were rival and inconsistent hypotheses that generated conflicting predictions in the important experimental situations. The inconsistencies are straightforward consequences of how classical particle ballistics and classical wave behavior were represented. Such a situation naturally leads one to suspect that neither model gets the whole story right, and that we ought to search for a more fundamental theory. In this case, such a suspicion turned out to be correct: in quantum electrodynamics, there is still a wave-particle duality, but it is not in the form of inconsistent models. Present-day physics is, however, plagued by a similar problem, perhaps of much greater importance, in that its (relativistic) model of gravitation and its (quantum) model of strong/weak/electromagnetic forces are inconsistent and yield nonsensical results in the special case of matter/energy that is both very small and very massive, such as conditions at the beginning of the universe or within a black hole. The search for a Grand Unified Theory is an effort to remedy this situation, and research programmes such as string theory are attempts to reduce classical and quantum theories to something yet more fundamental in which there is no longer any inconsistency.

Some theoretical physicists tend to be rather religious about their view that there must be a Grand Unified Theory. But in our present situation, what we can say is this: the very representational systems used for modeling gravitation and strong/weak/electromagnetic force yield inconsistencies in certain cases. These are *not* experimental discoveries about the Big Bang or black holes: they are derived solely from the theories. Each theory is, in its own right, extremely *apt*: each has been confirmed about as well as any claim in the history of science. And they are indeed consistent in the kinds of cases we generally observe. What we do not know is whether there is *some other* way of representing the phenomena that has equal or greater explanatory power but does not have the same sorts of formal snarls—in which case the problem turns out to be *merely* an artifact of a particular combination of representational systems—or whether any alternative account we come up with might face similar complications. Even this latter situation need not imply that the *world* would be inconsistent. Indeed, I tend to think that the sentence "The world is inconsistent" involves a grammatical or category error: *consistency* is something that applies to combinations of *representations*, not to objects or worlds. What it would mean is that the human mind can do no better than two separate representational systems for different phenomena, each of them very apt, which yield inconsistent results in boundary cases. What a Grand Unified

Theory would show, among other things, is that there is a single representational system available to human minds in which gravitation and strong/weak/electromagnetic force can all be consistently modeled in all known conditions simultaneously.

The long and the short of it is that scientific modeling is not simply a reflection of the world in a mirror of nature (Rorty 1979), but is always a representation of some features of it in the framework of some particular representational system. A number of philosophers of science have urged, additionally, that our methods of observation and intervention—both those natively endowed in our perceptual systems and those requiring instrumentation and laboratory procedures—also shape the kinds of models we employ and how they relate to the world. Thus Nancy Cartwright (1999) has argued, for example, that nature behaves in a lawful way only when regimented into careful arrangements called "nomological machines,"[15] and Ian Hacking (1983) has claimed that our experimental procedures create phenomena. I am in sympathy with this Pragmatist strand as well, though I think that pragmatism and cognitivism blend into one another without clear boundaries. On the one hand, we might think of the laboratory environment as part of our extended cognitive phenotype. On the other hand, we might view, not only science, but even the cognition of children and animals in Pragmatist terms, as elements in solving the problem of how to interact with the world adaptively. (As a pluralist nonreductionist, I have no urge to try to reduce pragmatic concerns to cognitive ones or vice versa. Each of these, like each apt scientific model, gives us an apt understanding of a limited slice of how things are.)

8

The Scope and Plausibility of Cognitive Pluralist Epistemology

Although I have presented the initial exposition of Cognitive Pluralism in the context of pluralist philosophy of science, I do not see it as fundamentally or exclusively a view *about the sciences*. Rather, I wish to make a case that most of the principal features of scientific modeling that were discussed in chapter 7 are in fact features of human cognition generally and reflect what we might plausibly view as deep "design principles of human cognitive architecture" (and indeed of principles of cognitive architecture shared with other species). On this view, science is an especially regimented and (hence) powerful extension of a more basic and widespread cognitive activity: *creating and utilizing mental models* of various corners of the world. (Though the emphasis here is on *cognitive* factors, this is not meant to preclude a role for other factors that might also shape models in a direction that leads to a plurality of models. For example, there are, to be sure, aspects of science that are necessarily social as well as cognitive. But (a) this is probably true of many other models humans employ as well, and (b) the end products of theories, laws, and models are the sorts of things that can be understood by individual minds, and hence are cognitive *among other things*.)

The goals of this chapter are therefore twofold. The first is to make a case that crucial features of scientific modeling, identified in chapter 7 as potentially contributing to an abiding theory pluralism in the sciences, are in fact quite generic features of cognition. In short, we understand the world through deploying a number of special-purpose, idealized, and partial models of aspects of the world, each of which is tuned to the demands of particular pragmatic contexts and employs a particular representational system that is suited to its problem

domain. Of course, one feature found in many scientific models—explicit mathematization—is a specialized feature not found in most other mental models, and probably is totally absent from the models of the world employed by nonhuman animals. But the mathematical character of scientific models, particularly in physics, is not so much definitive of their status as *models* as it is an indicator of the exacting *regimentation* required for some types of models.

Second, I make a case for the thesis that having a mind that interacts with the world through a plurality of special-purpose, idealized models is a plausible "design principle" for human and animal minds. On the one hand, this view is consonant with several themes in the sciences of cognition, such as modularity and domain-specific representation and reasoning. On the other hand, it is both an *adaptive* architecture and one that is plausible on evolutionary grounds.

8.1. Cognitive Division of Labor and Three Notions of "Modularity"

The notion of "modularity" came into vogue in the 1980s. One way of formulating the question of whether the mind is "modular" was posed in terms of the comparison of the mind to a digital computer. Both Turing and von Neumann computing architectures operate through a single central processing unit (CPU) that sequentially performs operations on symbols stored in memory on the basis of a combination of hardwired architecture and stored programs. Production-model computers are generally built on such an architecture, but they also generally involve additional circuits dedicated to particular functions, such as arithmetic operations and interfaces with other devices (e.g., a USB port). Such special-purpose "modular" circuits are generally hardwired rather than driven by a stored program, as the set of operations they perform is stable, task-specific, and is more efficiently carried out by hardwired circuits than by execution of a program.[1]

For those who viewed mind as a computer, it was natural to ask whether its structure is one that employs a number of special-purpose modules, or one that performs many or all of its functions through something analogous to a CPU and operating system. But while this question may have been motivated by the metaphor of the mind-as-computer or brain-as-computer, it is really separable from its metaphorical origins. Even if mind and brain do not share some of the features characteristic of Turing or von Neumann computation, such as sequential rather than parallel processing, one may still ask whether particular cognitive functions are performed by "special-purpose units."

The word 'modularity' is, however, one of those words that is used in a confusing variety of ways, often without explicit definition. There is a broad usage, often found in the empirical literature, in which 'modularity' signals only that the mind or brain has special-purpose *methods* for performing particular tasks, irrespective of *how* it executes them. (There is a similar ambiguity in computer science, in which special-purpose reusable sections of program code

are also referred to as "modules." There, one might distinguish between "hardware modules" and "code modules.") In this usage, "modularity" signals a kind of *thesis of cognitive division of labor.*

However, the word is also used in other ways that are stronger and more restrictive. For example, it is often used to signal the thesis that there are *dedicated brain areas* assigned to particular cognitive tasks—for example, layers of tissue in areas of the visual cortex that have functions like detecting boundaries. This is a thesis not only about *cognitive* division of labor, but also about *cerebral* division of labor. To avoid such similar-sounding labels, however, I shall speak of this anatomical notion of "modularity" as a thesis about *localization* of cognitive function in particular areas of the brain.

Pylyshyn (1999) and Fodor (1983) place a separate set of restrictions on "modularity," constraints not so much on *where* the processing is done (i.e., locally in specialized brain tissue rather than globally) but *how* it is done. In particular, "modules" in Pylyshyn and Fodor's sense must be "informationally encapsulated" and "cognitively impenetrable." The basic idea behind "encapsulation" is that the processing within a module is walled-off from information present elsewhere in the system.

> One can conceptualize a module as a special purpose computer with a proprietary data base, under the conditions that a) the operations that it performs have access only to the information in its database (together of course with specifications of currently impinging proximal stimulations) and b) at least some information that is available to at least some other cognitive processes is not available to the module. (Fodor 1983, 3)

A circuit is "cognitively impenetrable" in case its output "is largely insensitive to what the perceiver presumes or desires" (Fodor 1983, 68). For example, familiar visual illusions do not disappear just because the perceiver knows them to be illusory. This amounts to an informational insensitivity to at least certain types of information from other parts of the brain (and from the conscious mind, whatever its relationship to the brain). However, both Pylyshyn and Fodor allow that processes they consider to be "modular," such as Marr's (1982) "early vision," may be distributed over multiple neural areas and layers and involve feedback processing within the circuit. Their notion of "modularity" does not require anatomical localization.

Fodor also takes the view that "modularity" is not an all-or-nothing affair, but admits of degrees. For example, a "module" might be highly restricted in what sorts of information from other parts of the brain it is sensitive to without being insensitive to *all* such information. The contrast he wishes to make is with what he calls "central" cognition, such as that required for reasoning, which he claims requires global access to information. Fodor's view is that some cognitive functions, particularly those of early sensory processing, and perhaps a Chomskian "language area," are "modular" (i.e., encapsulated and cognitively impenetrable), but that most higher cognition is nonmodular (Fodor 1983).

My chief interest, by contrast, is with the broadest notion of "modularity," which I am calling *cognitive division of labor*. I wish to argue that

1. Cognitive division of labor is a deep and ecumenical design principle in human and animal brains.
2. At least one form this takes is in the creation of functional processes in the brain that model aspects of the organism, its environment, and relations between the two.
3. This division of labor can be realized in a number of ways, including but not limited to the localization of cognitive function in neural areas and layers.
4. We may see the extension of cognitive division of labor, from the proliferation of special-purpose *areas* to neurally distributed (though perhaps not fully global) special-purpose acquired *abilities* (through an intervening step of "redeployment" of ensembles of special-purpose areas to be explained later), as an instance of a *single design principle that is implemented in different mechanisms and on different timescales*: through gene selection resulting in special-purpose brain *areas*; through *redeployment* of existing areas in new functional *configurations*; and through *learning* in neural networks optimized to form pragmatically driven, interest- and organism-specific, idealized *models*.
5. Scientific models are special cases of these latter capacities.

Although it is crucial to my case that cognitive division of labor *need not* be confined to localized brain areas that are completely encapsulated, the discovery of such areas lends partial support to my thesis. Cognitive division of labor does not entail anatomical localization of function, but localization does entail at least a limited amount of cognitive division of labor. And given that even Fodor views encapsulation and cognitive impenetrability as admitting of degrees, we may for now leave it an open question how much of these features a given cognitive process involves (though we may signal in advance that there are good reasons why a division of labor accomplished through learning rather than special-purpose tissue would generally be less encapsulated and more cognitively penetrable).

8.2. Localization of Cognitive Functions in the Brain

There is now a great deal of evidence to support the view that many cognitive functions are anatomically *localized*: that is, that significant numbers of mental capacities are rooted in special-purpose mechanisms performed by dedicated bits of neural tissue. The first direct evidence for this thesis arose through trauma studies like those pioneered by Broca in the nineteenth century, in which it was observed that patients with highly localized brain injuries characteristically experience loss of narrowly defined psychological capacities as well, and that there are many robust correlations between brain areas and psychological functions.[2] Post-mortem examination (the only technique available in Broca's

day) of the brains of patients who had suffered strokes or head injuries revealed correlations between the areas that were damaged and the loss of particular cognitive functions, such as speech production or comprehension, impulse control, and recognition of human faces.

In recent years such results have received significant confirmation and been developed in much greater detail through brain imaging, experiments on animal models, and single-neuron sampling in patients undergoing brain surgery. Mapping the brain is proceeding at an amazing pace and is rewriting our understanding of the unities and disunities of ordinary cognition. To take only the case of vision, we intuitively think of "seeing" as a single mental process—Hume suggested the metaphor of understanding it as a drama played out on a single imagistic stage, for example. Yet information from the retina is quickly split into three separate pathways for boundary/shape, color, and motion, which are registered in different layers of cells in the LGN and separate regions of the visual cortex (Felleman and van Essen 1991).[3] All three streams of information pass through the LGN, V1, and V2, albeit through different layers of cells. V1 and V2 perform a great deal of the work in extracting information about form. Information about motion is passed on from V1 and V2 to V3 and V5, and information about color to V4. From the visual cortex, information is further split into a "what" pathway (located ventrally in the temporal cortex) and a "where" pathway (located dorsally in the parietal area). The "what" pathway is involved in recognizing types of objects and seems even to include such exotica as an area (the fusiform gyrus) specifically devoted to the recognition of faces. The "where" pathway is involved in both locating objects in space and in interacting with them kinesthetically (see Figure 8.1).

Such cerebral division of labor seems to be the rule rather than the exception in the human brain. Moreover, it is not only perceptual systems that display this type of localization: at the level of gross anatomy, the brain is divided into a number of areas that play different roles in perception, cognition, and behavior. To give a very rough overview that obscures many details: the brain stem controls autonomic functions such as blood pressure, heart rate, and breathing; the cerebellum controls balance and posture; the midbrain regulates blood pressure, hunger, thirst, circadian rhythm, and emotions; the thalamus works together with the cortex in processes of perception and movement. Turning to the cortex and mapping broad functions like vision and motor control onto Brodmann areas, a standard mapping is represented in Figure 8.2 and Table 8.1.

Recent techniques of neural imaging, particularly functional magnetic resonance imaging (fMRI), have greatly increased the pace at which scientists have been able to study how the brain operates when performing particular cognitive tasks. This branch of study, called functional neuroanatomy, relies in large measure on fMRI images taken of subjects performing particular cognitive tasks and comparing them with control images. Because fMRI measures blood flow in different regions of the brain, it is assumed to provide an indirect measure of levels of neural activity, on the assumption that heightened neural activity in a region increases the local metabolic demand, which is met through

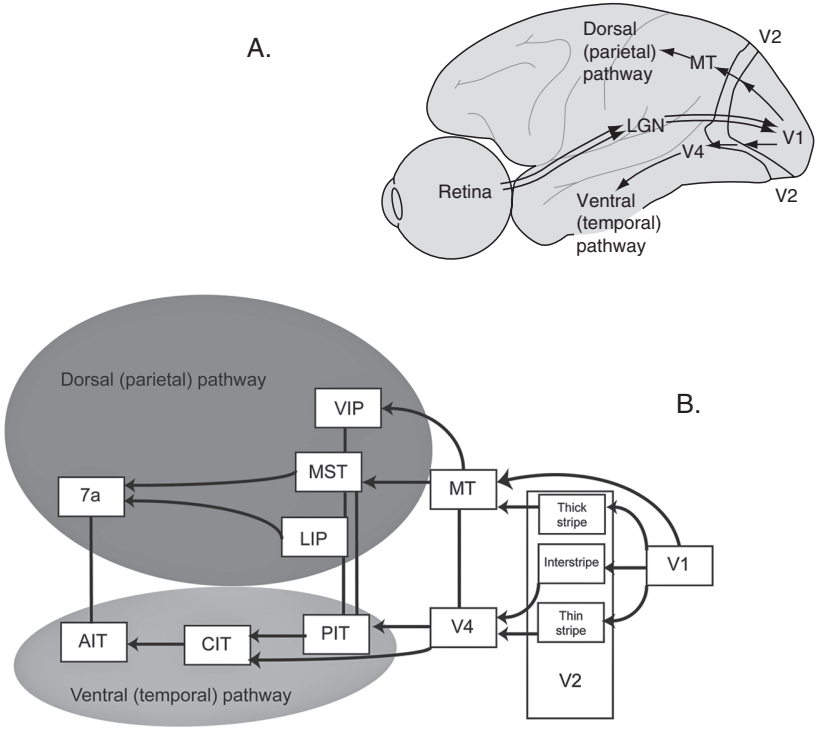


FIGURE 8.1. Ventral and dorsal visual pathways. (A) Anatomical and (B) schematic representations of visual processing in the ventral ("what") and dorsal ("where") pathways. The dorsal pathway passes into areas MT (middle temporal area), VIP (ventral intraparietal area), MST (medial superior temporal area), LIP (Lateral intraparietal area) and 7a. The ventral pathway passes into areas PIT (posterior inferior temporal area), CIT (central inferior temporal area) and AIT (anterior inferior temporal area). Note that this is a simplified diagram, in which many areas are omitted. Lateral connections between areas are indicated by lines without arrows. Adapted from *Principles of Neural Sciences*, edited by E. Kandel, J. H. Schwartz, and T. M. Jessell (New York: McGraw-Hill 2000), p. 550.

increased blood flow. Of course, blood is flowing throughout the brain all the time, and so the raw data of such images indicate activity across the entire brain. The pictures that are featured in textbooks, showing one or two areas of the brain "all lit up," are a product of *subtracting* base level activity, thus indicating which areas are *differentially* active in particular tasks. Researchers characteristically conclude that the areas that "light up" during a task are specially implicated in that task and are wont to report their results as showing that they have found a "memory area" or an "attention area."

Additionally, there are strong homologies between human brains and those of nearby species, where similar division of labor is found. (Indeed,

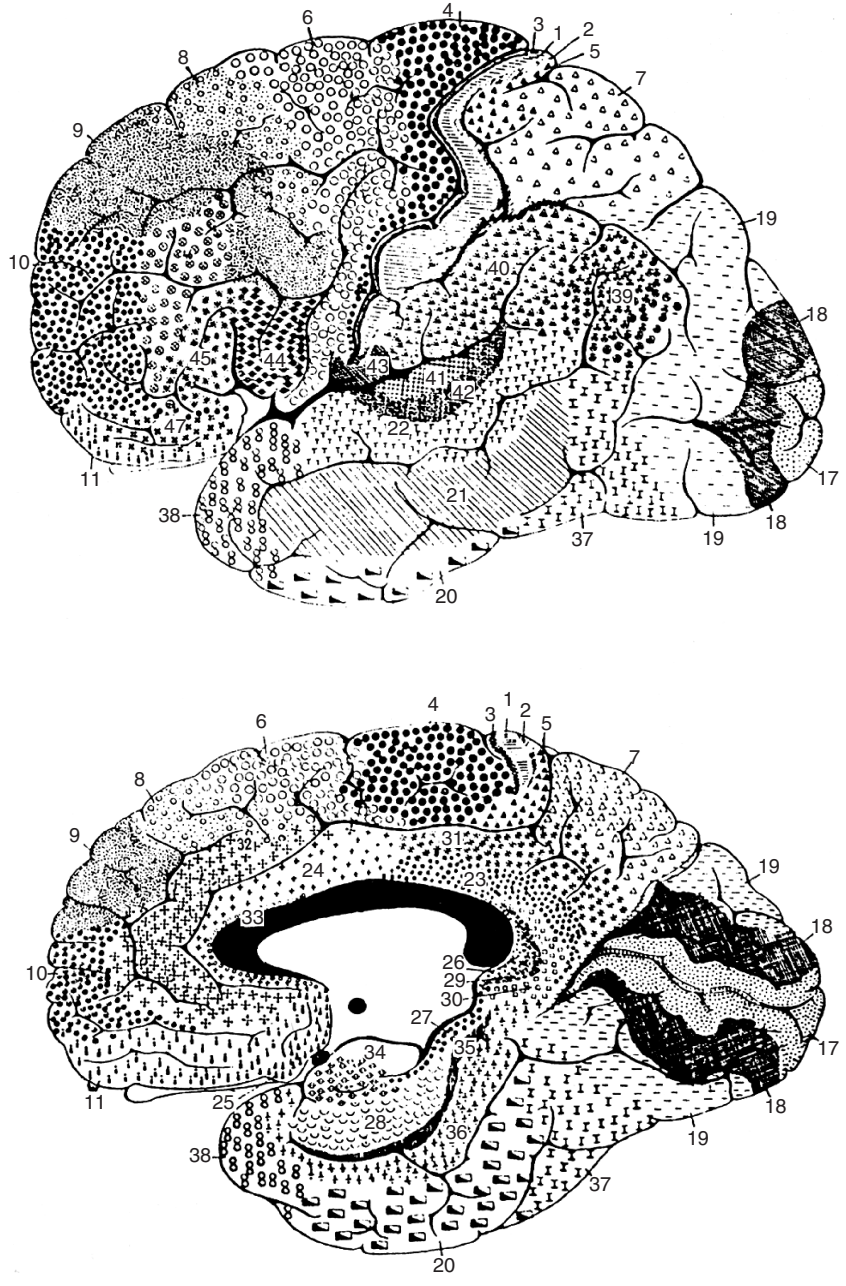


FIGURE 8.2. Brodmann's cytoarchitectonic map. Lateral and medial views of the human brain labeled by Brodmann area, from K. Brodmann, *Vergleichende Lokalisationslehre der Grosshirnrinde* (Leipzig: Barth, 1909).

TABLE 8.1. Brodmann Areas Organized by High-Level Functional Categorization

Function	Brodmann area
Vision	
Primary	17
Secondary	18, 19, 20, 21, 37
Audition	
Primary	41
Secondary	22, 42
Body Sensation	
Primary	1, 2, 3
Secondary	5, 7
Tertiary Sensation	7, 22, 37, 39, 40
Motor	
Primary	4
Secondary	6
Eye Movement	8
Speech	44
Tertiary Motor	9, 10, 11, 45, 46, 47

much research mapping the brain is done on animal models, such as cats and macaque monkeys.) Such homologies suggest that the specialization of brain areas, or distinctive pathways of processing through multiple areas, is a basic design feature of at least mammalian cognitive architecture. Indeed, when we look at more distant animal relatives, such as flies and bees, which lack the sophisticated structure of the mammalian cortex, we find a similar division of labor, in spite of considerable differences in brain anatomy (Paulk and Gronenberg 2005).

8.3. Complications

The story I have told so far is a somewhat sanitized and expurgated version, however. While there is little question that there are cognitive functions for which particular brain areas are very important in normal human brains, there are also several issues that stand in the way of a thoroughgoing "localist" hypothesis. First, there are also abilities that seem to resist any neat localization. Second, many brain areas exhibit a significant degree of *equipotentiality* or *plasticity*: the ability of parts of the brain to take on the functions normally performed by other parts of the brain when the latter are damaged or fall out of use. Third, meta-analysis of studies in functional neuroanatomy based on fMRI studies reveals that these seem to indicate a many-to-many relationship between cognitive functions and neural areas, rather than the one-to-one mapping that talk of "localization" might suggest.

8.3.1. Nonlocalized Functions

Recall that Fodor claimed that, although some neural systems (particularly early perceptual processing and a few more sophisticated functions such as language acquisition) are likely to be highly modular, the same cannot be said of many higher cognitive functions, such as reasoning. Recall that Fodor's notion of "modularity" was formulated mainly in terms of information access and not localization, and so his claims do not directly bear upon the scope of the localization thesis. But in fact similar claims could plausibly be made for localization as well. In particular, the regions of the human neocortex that are dramatically larger than those of any other species—and which are thus assumed to be the seat of distinctively human abilities such as reasoning—do not seem to subdivide in the same ways that evolutionarily older areas of the brain do. Indeed, even the "language areas" such as the Broca and Wernicke areas, long stock examples of localized functions, differ in location across subjects far more than areas like V1.[4]

A second problem is presented by cognitive skills that are plainly learned. The ability to play chess is a cognitive skill. But it is implausible that there is a "chess area" in the brain. And even if capacities to play chess *are* subserved by local groupings of cells in individual brains, it is unlikely in the extreme that such an ability is nativistic. Chess is just too recent an invention to have given processes of genetic variation and selection a chance to operate.

Another problem for localization is found at a smaller scale. It was at one time assumed that individual concepts like GRANDMOTHER must have a local neural basis, say, in the form of a single cell or group of cells whose activation underwrites thought using the concept. This, however, does not appear to be the case. For one thing, if it *were* the case, we should expect that routine cell death (say, after a night of heavy drinking at a fraternity party) would cause a more or less random loss of concepts throughout the human life span. But this does not seem to happen, though there are degenerative conditions such as forms of dementia that result in progressive and widespread loss of words and concepts. This suggests that, unlike simple connectionist models (e.g., Gorman and Sejnowki 1988), which tend to model concepts with single "output" nodes, the neural basis of concepts in the human brain must take some other form, such as a distributed activation pattern across a field of cells that is resilient in the face of the death of individual neurons (cf. Grossberg 1987).

At this stage of the development of the sciences of the mind, it is probably wise not to draw conclusions from such observations too quickly. The lack of results of a particular type—for example, failure to find a "neural correlate" of some particular function—may simply mean that we do not yet know where or how to look for it, or that we are framing our questions in the wrong way. For example, if we do not find "localizations" of particular concepts in the form of something on the order of "GRANDMOTHER cells," perhaps this is because individual concepts are not analogous to functional processes like the processing of color or shape, either. It may be that, in the case of concepts, what is

comparable to the functions of visual areas is something like a "distinction engine" or "discrimination engine," ranging over many concepts, much as the visual areas are able to represent many distinct visual patterns. Some of the kinds of models employed by connectionists interested in concepts suggest a "representation" of concepts that is *distributed within* the network (say, in the form of a distinctive spatial pattern of activation levels or activation ratios, or in the global pattern of connection weights).

But the cells comprising such a network are most plausibly interpreted as being a small subset of the cells in the entire brain. In humans as opposed to computer models, such a network might well be "localized" in a set of cells in the neocortex, even if (a) the cells are not anatomically adjacent to one another, and (b) their partition and function is not nativistic, but a function of the training of the network through perception and learning. Here it is useful to distinguish several "localist" (or "modularist") theses that are importantly different from one another:

- *Nonglobalism*: A function F is subserved by a set of neurons that is a (small) proper subset of the whole brain. (This set might be spatially distributed and partitioned as a functional unit through learning.)
- *Anatomical Localization*: A function F is subserved nonglobally by a set of neurons that corresponds to a unit identifiable on anatomical grounds—for example, a layer of cells within a Brodmann area.
- *Nativism*: A function F is subserved nonglobally by a set of neurons that are nativistically determined (or strongly biased) toward performing F.

Nonglobalism is a necessary but not a sufficient condition for anatomical localization and nativism. And a function can be either anatomically localized or nativistic without being both, as we shall see in the next two subsections.

In the case of things that are learned, such as most of our concepts, and acquired skills such as chess playing, we probably should not *expect* them to be nativistic, nor should we expect anatomical localization beyond the use of broad areas of the brain for concepts. If the function of the human (and probably mammalian) conceptual system is to be a kind of *discrimination engine* that acquires specific discriminative and recognitional abilities through experience, we should not expect the brain to come prestocked with particular concepts, or at least not the kinds of concepts that are learned. An efficient architecture for learning of this sort is a network architecture (or perhaps one that involves the ability to create *multiple* semantic networks) whose "shape" is driven by learning rather than by anatomy. Anatomy may determine what cells are available to be utilized in such learning, and by the algorithmic shape of the learning process, but the resulting functional architecture (including its bases in both connection strengths and growth and paring of dendritic connections) will be highly dependent on learning history and will vary from individual to individual. Moreover, individual brains will differ in the number and density of cells involved in such a process. Normal human brains may be similar at the

level of Brodmann areas, yet differ widely in the number of cells and topology of neural connections within those areas.

It is thus possible that *all* cognitive functions are nonglobal, even if some of them are nonnativistic (at least in their details, such as *what* concepts an individual possesses) and not anatomically localized at a fine-grained level.

8.3.2. Equipotentiality and Plasticity

A second challenge to nativism and anatomical localization stems from the fact that the brain is a highly flexible organ. Often, when particular cognitive abilities are lost due to strokes or injuries to the brain, they are to some extent regained, through a process in which other parts of the brain take over the functionality of the damaged tissue. Likewise, if an intact area of the brain can no longer perform its normal function (say, the visual cortex is deprived of visual inputs through blindness), its cells may be co-opted to perform other functions.

Mriganka Sur and his MIT colleagues, for example, have performed numerous experiments on the auditory and visual connections of the brains of ferrets. In ferrets, as in humans, auditory signals pass through the thalamus before reaching the auditory cortex. But whereas these connections are present at birth in humans, they develop after birth in ferrets, and hence interventions are possible in neonatal ferrets. Sur found that if he cut the connections of the auditory stream to the thalamus, the optic nerve would grow connections to both "visual" and "auditory" areas of the cortex (Sur, Garraghty, and Roe 1988). More surprising still, in these ferrets the auditory cortex would develop the "pinwheel" organization of cells normally found in the visual, but not the auditory, cortex. These formations were not as numerous or as orderly as those in the visual cortex, but were nonetheless distinctive anomalies not found in normal auditory areas (Sharma, Angelucci, and Sur 2000). The portions of the ferrets' brains that would normally have served as an auditory cortex instead became (functionally) a second visual area and developed a structure with features associated with a normal visual cortex. Sur's conclusion is that such brain regions exhibit a significant degree of developmental plasticity and acquire their "function" only through development and experience.

For a variety of reasons, it is not possible to perform the same experiments on humans. Apart from the moral issues that would be involved, human infants are born with the connections from eyes and ears to visual and auditory cortex already intact, and so the interventions would need to be performed on human fetuses. However, Sur's findings have resonances with long-time reports of increased acuity of other senses when one sensory modality is lost—for example, increased auditory and tactile sensitivity in the blind. Until recently, such reports were based largely on anecdote; but more recently, technologies like fMRI have allowed researchers to investigate the neural activity of both blind patients and of experimental volunteers who are temporarily deprived of visual input. Perhaps the most intriguing experiments in this vein were performed by Dr. Alvaro Pascual-Leone of Beth Israel Deaconess

Hospital in Boston and his colleagues. Pascual-Leone blindfolded sighted experimental subjects for a period of five days and measured their brain activity as they performed various tasks over the course of that period. Over the course of the trial, the "visual" (i.e., occipital) cortex began to "light up" during the performance of tactile and auditory tasks. Pascual-Leone's conclusion based on these results was that this period of sensory deprivation seemed to be "sufficient to lead to recruitment of the primary visual cortex for tactile and auditory processing" (Pascual-Leone et al. 2005, 390).

The blindfold experiments are importantly different from Sur's experiments, in that they were performed on adult human subjects without brain injuries and the results were both rapid in onset (over the course of days) and highly transient (they disappeared after blindfolds were removed). This strongly suggests that, unlike Sur's ferrets, the subjects in the blindfold experiment did not undergo significant structural reorganizations in the brain.

> The speed of these functional changes is such that it is highly improbable that new cortical connections are established in these sighted individuals. Therefore, somatosensory and auditory connections to the occipital cortex must already be present and are unmasked under our experimental conditions. These could be cortico-cortical connections, linking Heschl gyrus or postcentral cortex and striate cortex directly, via cortical multisensory areas, through thalamic or other subcortical relay nuclei. Ultimately, the occipital cortex recruitment mechanisms in tactile processing in the blind and under blindfolded conditions are not likely to be identical. (Pascual-Leone et al. 2005, 14–15)

The fact that there are connective pathways from auditory and tactile transducers to the "visual" (i.e., occipital) cortex is itself highly interesting. It, and the experimental data for cortical plasticity, render it necessary to separate our terminology for referring to a region of brain *anatomy* (e.g., "occipital cortex") from our terminology for referring to it by its *function* (e.g., "visual cortex"). Whether the occipital cortex takes on visual functions is dependent on factors in development, and its continued performance of those functions is sensitive to injury and sensory deprivation.

What are the implications of such experiments for the localization theses? First, they do *not* provide evidence against nonglobalism. In Sur's ferrets and the subjects of the blindfold experiment, what is found is not *global* activation across the entire brain in all cognitive tasks studied, but rather a change in the mappings between structure and function. Indeed, what is so striking about these experiments is that specific unexpected areas "light up" in the performance of particular tasks.

They do, however, provide evidence against the view that the normal assignments of functions to anatomical areas is *strongly* nativistic, in the sense of being determined by genetics independent of development, injury, and experience, or even by the empirical measure of being universal within a

species. However, this problem is less serious than at first it might appear. Research in developmental systems in biology has shown that genetics rarely fully determines a phenotypic trait: most traits are "plastic," sometimes quite highly so. Harvard biologist Richard Lewontin is eloquent on this point in an interview with Werner Callebaut:

> Any textbook or popular lecture on genetics will say: "The gene is a self-reproducing unit that determines a particular trait in an organism." That description of genes as self-reproducing units which determine the organism contains two fundamental biological untruths: *The gene is not self-replicating* and *it does not determine anything*. I heard an eminent biologist at an important meeting of evolutionists say that if he had a large enough computer and could put the DNA sequence of an organism into the computer, the computer could "compute" the organism. Now that simply is not true. Organisms don't even compute themselves from their own DNA. The organism is the consequence of the unique interaction between what it has inherited and the environment in which it is developing (cf. Changeux 1985; Edelman 1988a, b), which is even more complex because the environment is itself changed in the consequence of the development of the organism. (Callebaut 1993, 261)

Moreover, philosophical discussions of "nativism" have distinguished several "degrees" of nativism, ranging from full determination to developmental biases (Ramsey and Stich 1990; Cowie 1999). The Sur and Pascual-Leone experiments are fully consistent with nativistic *biases*. Indeed, both researchers assume that normal development will characteristically result in the traditional pairing of structure and function.

The experimental findings are also consistent with a form of anatomical localism. In any organism studied, at any one time, there are particular regions of the brain that are implicated in the performance of visual, auditory, and tactile perception. The experiments do show that there is plasticity in *which* areas can do so, and *how many* may do so. In Sur's experiments, this is a difference between normal and "rewired" ferrets; in Pascual-Leone's, it is a difference in the performance of individual human brains over time. The plasticity demonstrated in these experiments shows that the function-to-structure mappings are not species-constant, nor necessarily even constant in a single organism over time. But they are also consistent with (and indeed assume) the thesis that at any one time, there are real function-to-structure mappings.

These experiments also have potential implications for my cognitive pluralist claim that organisms are likely to employ different representational systems for different cognitive tasks, with the system "chosen" (through evolution or learning) being driven in large measure by the informational constraints of the task, as well as the constraints set by biological facts about the organism. Sur's experiments require us to make an important clarification in this thesis. If the thesis were that anatomically typed areas are genetically

determined to have a particular architecture that supports a particular type of representational system, Sur's experiments would show that this is not universally true, as the ferret's "auditory" cortex develops some features distinctive of "visual" areas in rewired ferrets. But Cognitive Pluralism was never a claim about nativism. And Sur's results actually *strengthen* my case in unanticipated ways: if the rewired areas are coupled with visual tasks during the critical developmental period, the development of their "wiring" is influenced in directions that more optimally support visual tasks. The fact that the resulting neural wiring still has some "auditory" features suggests that there are some developmental biases that cannot fully be overcome by visual input in the critical period, but it still looks as though developmental and experiential factors play a large role in shaping the neural architecture into one that supports a task-specific representational system.

But if this is the case, might not the blindfold experiments support the opposite conclusion? In those cases, the occipital cortex is co-opted for tactile and auditory functions, but on a timescale too short for the explanation to be found in "rewiring." Indeed, the change seems to *increase* the acuity of the nonvisual senses. Isn't this in tension with the idea that the occipital cortex (i.e., of sighted adults) employs a representational system that is particularly suited to *visual* tasks? Not necessarily. In the blindfold experiments, we are dealing with subjects who already have normal tactile and auditory abilities, and so these experiments do not allow us to compare the performance of "visual" versus "auditory" areas in underwriting the capacities that are exercised. Rather, in these cases, whatever was previously present is *supplemented* by whatever the "co-opted" areas of the cortex contribute.[5] We don't know what sort of auditory sensory accuity occipital processing of auditory information, of the sort encountered in the blindfolded subjects, would offer without the "normal" auditory processing that is left intact.

8.3.3. The "Massive Redeployment Hypothesis"

A slightly different challenge to localism is presented by a meta-analysis of over one hundred fMRI studies by Michael Anderson (forthcoming). Each of the original studies investigated the neural correlates of the performance of some

TABLE 8.2. Number of Brain Regions Activated (out of 31), with Activations in Exactly the Number of Task Categories Listed, out of the Four Categories Surveyed

Activation type	Number of areas with activation in *exactly*			
	1 task category	2 task categories	3 task categories	4 task categories
Right lateral	3	4	11	11
Left lateral	2	3	8	15
Right medial	4	4	6	0
Left medial	6	3	4	2

Source: Reproduced from Michael Anderson, "The Massive Redeployment Hypothesis and the Functional Topography of the Brain," *Philosophical Psychology* (forthcoming). By permission of author.

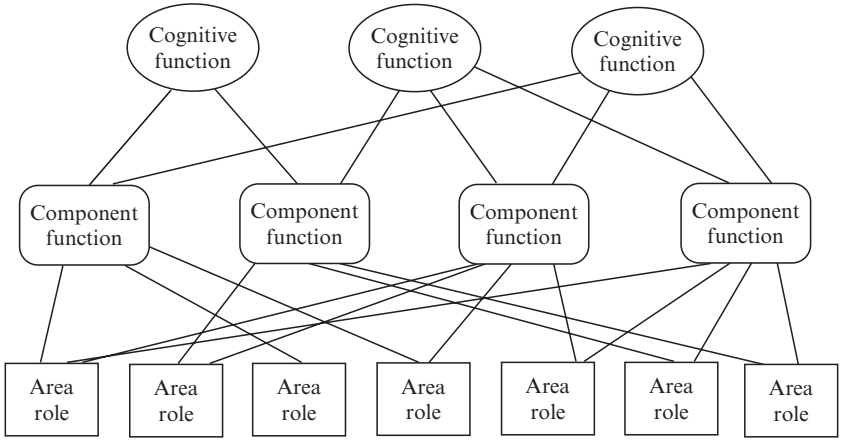


FIGURE 8.3. A three-tier architecture showing a many-to-many relationship between all levels. Reproduced from Anderson (forthcoming). By permission of author.

cognitive task and identified, through subtraction analysis, areas of the brain that were differentially active during the task. (Thirty-one Brodmann areas were represented across the original studies.) The original authors of the studies on which Anderson drew tended to reach the conclusion that the regions identified were therefore "memory regions," "attention regions," and so on, depending on the nature of the task they were studying. Anderson's analysis, however, revealed that (a) most regions studied in the experiments were utilized in multiple tasks, and indeed multiple *types* of tasks (e.g., attention and memory), and (b) most tasks involved multiple Brodmann areas (see Table 8.2).

Anderson's analysis shows that "cognitive tasks" do not stand in a one-to-one relation with Brodmann areas, but in a many-to-many relationship (see Figure 8.3). How should we explain such findings? Anderson's suggestion is that what we are seeing is the result of a "massive redeployment" of preexisting brain areas to obtain new functionality. He hypothesizes that, in evolutionary history, the brain areas thus redeployed were originally more truly modular units, but that evolution took on a new strategy for acquiring new functionality somewhere along the way. This new strategy consisted in redeploying existing functionality in ensembles of neural areas working together rather than waiting for gene selection to produce it by way of mutations producing new brain areas.

Anderson's explanatory hypothesis, of course, goes beyond the findings of his analysis of studies in functional neuroanatomy. Confirming it would require cross-species comparisons and a plausible model of evolutionary stages leading to the features found in the human brain. Nonetheless, his study presents an interesting challenge to localism, and massive redeployment represents an important possibility that is neither fully localist nor fully globalist in its lineaments.

Massive redeployment is compatible with nativism, though it does not require it. It implies nonglobalism, though it differs from a stricter sort of

localism that requires that the regions employed in a cognitive function must be anatomically contiguous. It also implies cognitive division of labor, while suggesting that there is a hierarchic structure to this division of labor. (Metaphorically, tasks are assigned to platoons, working groups, or committees rather than individuals.)

8.3.4. Summary of the Analysis of the Complications

There are indeed forms of the modularity hypothesis, localism, and nativism that are called seriously into question by the examples we have discussed in this section. However, none of the examples imperils what is essential for establishing that Cognitive Pluralism is a basic principle of cognitive architecture: namely, cognitive division of labor. Indeed, each of these results supports, rather than contradicts, this hypothesis, as all are at odds with globalism. They do, however, have implications for what can count as a viable hypothesis on the empirical question of *how* cognitive division of labor is achieved. Sur's explorations of cortical plasticity suggest that we ought not expect all cognitive division of labor to be determined by genetics, but we should be sensitive to developmental conditions as well. Pascual-Leone's experiments suggest that, even once a brain has gone through critical developmental periods and hence has been "wired" for a particular division of cognitive labor, it may still be possible to reallocate particular "jobs" to different areas. And Anderson's massive redeployment hypothesis suggests that some tasks may be carried out "by committee" rather than by individual anatomically segregated units.

All of this is quite congenial to Cognitive Pluralism as an empirical thesis about cognitive architecture, as is the implication that not all functions need be linked to anatomical areas in the same fashion. Indeed, in the next section I suggest that there is also a need to find room for cognitive models whose distinctive representational schemes are not directly rooted in brain physiology at all. This, moreover, is needed to accommodate concerns about examples like chess (and science), for which it is implausible that there are special brain modules.

8.4. The Binding Problem

We began the case for Cognitive Pluralism with examples from neuroscience, such as the existence of separate visual streams for color, form, and motion. But the existence of such separate processing streams, in itself, does not refute the idea of a Humean "theater" of the mind. Visual information may be divided into separate processing streams in the retina and LGN, but the information streams might be reintegrated at some later stage of processing downstream from the visual cortex. And indeed, we do not (except under rather unusual and unnerving circumstances) *experience* three separate visual worlds. Rather, we experience one world, with color, shape, and motion all mixed together. When we see the world, we do not go back and forth between

something resembling line drawings without color and something resembling Impressionist paintings without object borders. Rather, we see a world of colored objects with well-defined boundaries, often in motion. How does this come about?

This is one way of framing the "binding problem": there are separate representations in the brain of color, form, and motion. For example, if I am presented with a red triangle and a blue circle, there are parts of the brain (perhaps in V4) that indicate "red" and "blue," and other parts (perhaps in V1 and V2) that represent "triangle" and "circle." But because these are separate representations, in different parts of the brain, we need a story about "how the brain knows that 'red' goes with 'triangle' and 'blue' goes with 'circle.' " Why don't we instead link "red" with "circle," or with both "triangle" and "circle," or indeed with no shape at all?

The "natural" conjecture would be to suppose that there must be another "module" in the brain where the information is recombined into a more comprehensive representation of a visual scene, with color and shape both "painted in" (and in examples involving motion, that as well). But current evidence does not support this hypothesis. As Dennett (1991) has argued, the "theater" metaphor for the mind (or even for a more limited subset of mental phenomena, such as those involved in visual perception) appears to be a theoretical fiction. There *are* proposals for how binding is accomplished, but they do not involve the postulation of a "theater" module.

The most celebrated attempt to solve the binding problem is that proposed by the late Francis Crick (of DNA fame) and Christoph Koch (1990). Crick and Koch suggest that binding is achieved through a kind of phase locking between the cells in different areas: in our example, the cells indicating "red" fire in phase with the cells indicating "triangle" and out of phase with the cells indicating "circle" (and likewise, *mutatis mutandis*, for "blue" and "circle"). To be sure, this does not explain why there is an *experiential* or *phenomenological* unity associated with phase binding, any more than the firing of cells in V4 explains why there is a distinctive phenomenology to seeing red. What it does do is provide a possible mechanism for how binding of *information* in two processing streams is achieved, and one that does so without positing an area with a unified representation of form and color.

Such ideas have received philosophical attention before. Minsky (1985) proposes the metaphor of a "society of mind." Dennett (1991) speaks of "multiple drafts" of information bouncing around in a "pandemonium" of different modules. Both Minsky and Dennett share my cognitive pluralist sentiments in seeing the mind as being, in important ways, disunified. But my thesis—and I am not fully clear on where Minsky and Dennett stand on this question—is about a very special *kind* of disunity: namely, *representational* disunity. In the present case, there is no unified representation in the brain, housed in its own Brodmann area, that represents a scene with both color and shape. Rather, these different features are represented in different parts of the brain, each employing a representational scheme appropriate to its subject matter. The brain *does* in some sense "unify" these representations, but not

by building a more comprehensive representation in another area. Rather, they are (in terms of mechanism) linked through phase binding (on the Crick-Koch hypothesis) and (in terms of function) are used to *triangulate* objects to which both sets of properties belong. Denying that there is a unified representational system for form and color in the brain does not require us to deny that there are *other* forms of unity present: in the case at hand, the kind of unity involved in constituting something as a single object with multiple properties, though one might add to this other types of unity, such as cross-modal identifications involving sound, sight, and touch, constituting successive experiences as experiences of the selfsame object, the transcendental unity of apperception, or personal identity. A cognitive pluralist does not need to deny the reality of these other forms of unity in denying unities of representation.

8.5. Domain-Specific Knowledge and Reasoning

Thus far, we have concentrated our attention on evidence for Cognitive Pluralism stemming from studies of the brain. An additional source of evidence linking scientific plurality to more ecumenical processes of cognition is to be found in studies of the domain-specificity of nonformal understanding and reasoning. Notions of "domain-specificity" have played important roles in linguistics, developmental psychology, and artificial intelligence. Chomsky (1965, 1966), for example, argued that we have special mental abilities for learning a grammar on the basis of a poverty of the stimulus argument: children learning a language converge on a set of grammatical assumptions that are radically underdetermined by the data with which they have been presented, and hence there must be something about the mind that is designed so as to constrain the grammatical search space. This need not be hardwired, but can be emergent from developmental biases plus social interaction in critical periods (Ramsey and Stich 1990; Cowie 1999), though there do seem to be "language areas" of the brain. It has likewise been suggested that there seem to be innate propensities to represent medium-size objects in particular ways ("folk physics"), to represent social relations, and to have a "theory of other minds."

In the developmental literature, the notion of "domain-specificity" is often linked to the idea of innateness. Yet this linkage strikes me as spurious. The knowledge and skills involved in playing chess or understanding and applying scientific models would also seem to be domain-specific, even though it is dubious in the extreme that such abilities are innate. Another tradition of thinking about domain-specific knowledge, artificial intelligence, has given more attention to domains of this sort. In particular, "knowledge representation" was one of the major focuses of second-generation AI.

Whereas the first generation of artificial intelligence research concentrated on the automation of formal operations and aimed at producing machines and programs that could prove theorems in logic and mathematics, the 1960s and 1970s saw attempts to model substantive human knowledge. Terry Winograd's

(1972) SHRDLU, for example, was a model of representation and reasoning about three-dimensional objects in a simplified block-world. Roger Schank's group (Schank and Abelson 1977) set out to model knowledge of particular social contexts, such as behavior at a restaurant. Whereas the design of theorem provers was a relatively straightforward task of implementing known formal algorithms and heuristics in a machine, modeling substantive knowledge domains required a great deal of new thought in order to understand the largely implicit and unconscious "commonsense" knowledge people employ in reasoning about, say, blocks or restaurant behavior, or even games that can be given definitive formal representations, like tic-tac-toe or chess. More than thirty years ago, Marvin Minsky (1974, n.p.) wrote:

> It seems to me that the ingredients of most theories both in Artificial Intelligence and in Psychology have been on the whole too minute, local, and unstructured to account—either practically or phenomenologically—for the effectiveness of common-sense thought. The "chunks" of reasoning, language, memory, and "perception" ought to be larger and more structured; their factual and procedural contents must be more intimately connected in order to explain the apparent power and speed of mental activities.

As suggested by this quote from Minsky, theorists of "knowledge representation" tended to see their work as closely connected to cognitive psychology. For example, SHRDLU's abilities to navigate its block world might be seen as a kind of simulation of aspects of "folk physics." Schank's group modeled the structures involved in understanding stereotyped social situations. Other researchers modeled semantic relationships at a lexical level in "semantic networks" (Quillian 1968; Norman and Rumelhart 1975). A number of researchers in the 1970s undertook the modeling of sentential-level inference. Abelson's group developed models of "implication molecules" (Abelson and Reich 1969) and "conceptual dependency analysis" (Abelson 1973) in cognition generally, and Colby (1975; a psychiatrist by training) modeled the peculiar inference patterns of a paranoid personality in PARRY. Already in the 1950s, Newell and Simon (1956, 1963) had developed the GPS (General Problem Solver) model of means-ends analysis. And Hayes-Roth (1980) developed a model of opportunistic planning.

The trend in this research was to move beyond attempts to find a single general model of reasoning and knowledge representation and concentrate instead on special-purpose models that were applicable to particular problem domains. Again quoting from Minsky, this time speaking retrospectively in 2004:

> To build a machine that has "common sense" was once a principal goal in the field of artificial intelligence. But most researchers in recent years have retreated from that ambitious aim. Instead, each developed some special technique that could deal with some class of problem well, but does poorly at almost everything else. We are convinced, however, that no one such method will ever turn out to be "best," and that instead, the powerful AI systems of the future will use

> a diverse array of resources that, together, will deal with a great range of problems. To build a machine that's resourceful enough to have humanlike common sense, we must develop ways to combine the advantages of multiple methods to represent knowledge, multiple ways to make inferences, and multiple ways to learn. (Minsky, Singh, and Sloman, 2004, 113)

This line of thought was also the core of Minsky's (1985) book *The Society of Mind*.

Here we see two important strands of Cognitive Pluralism at work: the thesis that the mind employs multiple strategies for different problems, and the thesis that these strategies employ representational schemes and methods that are specifically suited to their problem domains. In such a system, logical goals like global consistency are not the highest priority. Indeed, consistency may stand in the way of having more local or regional techniques that are useful in addressing distinct real-world situations. Commenting on another researcher's attempt to model commonsense knowledge, Lenat's Cyc, in an interview, Minsky is quoted as saying:

> They've made it consistent, so it actually doesn't know much. Should a whale be considered a mammal or a fish? Whales have many fish-like characteristics, so most people are surprised when they hear it's a mammal. But the real answer is, it should be both. A common-sense database shouldn't necessarily be logically consistent. Lenat finally realized that they should restructure Cyc by providing for the different contexts in which a question may come up. But the database was originally structured to make things very logical, and its language is predicate calculus. Our hope is to make the Open Mind system use natural language—which is of course full of ambiguities, but ambiguities are both good and bad. (Roush 2006)

Minsky's point here is consonant with my suggestion that the very features that make individual models adaptively suited to particular problems may render them unsuited to combination into a globally consistent system. Consistency may be an important goal in logic and mathematics, but it can lead us astray in understanding the mind. The mind seems to have been designed so as to be able to generate ways of understanding the world that are individually useful in solving particular pragmatically defined problems, and global consistency does not seem to be a design principle of our cognitive architecture. Not only is consistency the "hobgoblin of small minds," but a design for cognitive architecture would seem to achieve consistency only at the expense of adaptive power. And models of the mind that aim at a consistent system of representations and reasoning techniques will tend to be correspondingly psychologically unrealistic.

With a few exceptions, the kinds of mental abilities studied by knowledge modelers are largely not of the sort that are plausibly supposed to be innate. AI

models tend also to be neutral on the question of whether these learned models are anatomically local. The evidence for distinct learned, domain-specific models is independent of the question of how these are realized in the brain. Such models, however, clearly must be nonholistic. But the basic units of understanding in such models are not truly atomistic either. The knowledge of how to play chess, or to behave in a restaurant, is characterized by a fairly tightly knit set of beliefs and skills that are relatively independent of those employed in other domains. Take something one learns a ways into one's learning of chess, such as castling. Learning about castling changes one's understanding of the game of chess, and this change in understanding ramifies in such a way that the possibility-space of a game of chess is substantially altered. So acquisition of an understanding of castling is not epistemically *atomic*: it has intrinsic connections with other chess-knowledge. But it is not epistemically *holistic* either: learning to castle does *not* have constitutive effects upon, say, my "folk physics," my theoretical physics, my understanding of restaurant etiquette, and so on. Of course, my chess-knowledge is not wholly walled off from other knowledge either: I know how to *identify* physical objects as knights and queens, and I know how to move pieces on a tabletop chess set or a game of chess on my computer. Indeed, learning that these pieces were nailed down to the table *would* affect my beliefs about whether I could, in fact, move this (token) knight, even though it would not affect my knowledge of the rules of the *game* of chess.

Schank and others suggested that we view these relatively local knowledge domains in terms of information structures, given names like "frames" and "scripts." Knowledge modelers of the 1970s tended to assume that these could be well-represented in the rule-and-representation architecture generally used in computer programming. This assumption is arguably problematic, as argued seminally by Hubert Dreyfus (1979) and also by Terry Winograd (Winograd and Flores 1986), the designer of SHRDLU, on the grounds that expert knowledge requires both background kinesthetic and perceptual skills that are implicit and not rule-like, and also domain-specific expert skills that cannot be formalized. Even to the extent that it has proved possible to simulate human know-how in particular frames through algorithmic means, our ability to know how and when to *shift* the operative frame has proven less susceptible to algorithmic approaches. Yet this does not show that there *are no* "frames" in the sense of relatively autonomous domains of acquired knowledge and skill; it merely shows that these may not be realized through rules and representations. Indeed, I regard both frames (in the sense of local domains of knowledge and skill), and the problem of how one knows how and when to shift the operative frame, as powerful *prima facie* data for cognitive psychology, and that they represent an important research agenda. For example, can neural network architectures provide better models of how knowledge domains are acquired and how we reason in these domains (Winograd and Flores 1986)? Is cross-domain reasoning enabled by, and perhaps only by, possession of a language (Bermudez 2003; Sterelny 2003)?

8.6. Cognitive Pluralism and Domain-Specificity

Modelers of knowledge domains have for several decades been pursuing projects that are consonant with Cognitive Pluralism. The guiding assumption is that the types of knowledge and reasoning we employ in specific contexts like playing chess or ordering in a restaurant constitute relatively autonomous domains of representation, knowledge, reasoning, and action. They are "relatively autonomous" in the sense that (a) they employ proprietary representational systems, with their own domains, structures, and rules of inference, and (b) we can acquire, utilize, and revise our ways of relating to these domains in ways that have comparatively few ramifications in other areas of understanding, knowledge, and reasoning. *Pace* Quine, acquisition of chess concepts or changes in my understanding of the rules of chess would *not* have consequences for *all* the rest of my knowledge and concepts. Such understanding is not holistic. But neither is it atomistic: individual chess concepts and individual rules of the game have no meaning at all except in the broader context of an understanding of the game as a whole. And changes in my understanding of the concepts and rules *do* have constitutive implications for everything else *within* the game. Everything I think I know about chess strategy, for example, would have to be completely rethought if I suddenly discovered that I had mislearned rules of movement from the outset, and that knights move diagonally, like bishops.

Here we need a term that stands between 'holism' and 'atomism' and between 'globalism' and 'localism'. We might use the term '*regionalism*' to capture this intermediate scope of knowledge domains. (I am sometimes tempted to call it 'halfism', in contrast with holism; but my attractions to *that* pun are offset by the recognition that one could also pun upon 'halfist theory' in ways that I might later regret.)

The autonomy of regional knowledge domains, however, is not complete. Likewise, the autonomy of scientific domains is not complete, the encapsulation and cognitive impenetrability of "modular" systems admits of degrees, and the disunity of visual streams does not mean that they cannot influence one another or contribute jointly to other processes, such as depth perception. While the cognitive pluralist, by definition, resists the assumption that *representational unity is the norm*, she need not resist the idea that there are important, if partial, *connections* between models of regional domains, nor deny that we can deploy them in tandem to get a better understanding than we can from any of them taken individually. We have already seen this in the examples taken from the sciences: while gravity and electromagnetism are independent and not reducible to a common and more basic theory, we can combine their contributions through vector algebra for purposes of good-enough predictions.

My suggestion is that this is not merely a matter of analogy. Scientific models are a special case of mental models generally. They are a distinctive sort of model, in that they are particularly regimented, both in their mathematical form and in their methodologies. But they share with other mental models the fact that they are partial, idealized, and require particular representational

systems. The barriers to their integration are likewise a special case of the barriers to integration of mental models generally.

At least *some* of the causes of disunity among more mundane mental models are the same as those identified for the disunity of the sciences: their idealized character and their use of proprietary representational systems. A model of the game of chess (whether in a human mind or in a computer simulation) requires a number of distinctive representational features: a type of board with a distinctive geometry and topology, an "ontology" of pieces with distinctive properties of movement, rules of capture, starting positions, and checkmate and stalemate conditions that end the game. Such rules are generative, in that they define both a space of possible games and a space of possible strategies. Expert (i.e., master-level) understanding of the game may also involve skills that cannot be reduced to rules; but these, too, are constrained by the rules of the game. (Even a grand master cannot move a rook diagonally or escape from checkmate.)

One's understanding of a *particular* game, understood through one's chess model, but using a particular board and set of pieces, is idealized. Most palpably, it idealizes away from things like the physical nature of the board and pieces. Sometimes, these idealizations may matter in the playing of the game. They may affect how the game is in fact played. (For example, if I believe a particular piece is covered with a corrosive substance, I may refrain from touching it, thus effectively removing it from my strategy.) And they may disrupt the game altogether. (For example, if a badly designed chess program does not allow certain legal moves, I cannot in fact make those moves. Or if a piece is glued to the board, or a section of the board is broken off, the game cannot continue in the normal way.) My understanding of the physical situation—for example, that a piece is glued down—can influence how I play the game, or cause me to conclude that the game must be abandoned. But it does so in a fashion not describable within my model of chess itself. The game does not tell me how to proceed if a piece cannot be physically moved. I must improvise outside the rules, say, by moving to another board, or by suggesting to my opponent a modification of the rules (e.g., that neither of us will be allowed to move our rooks in order to level the playing field) that will allow us to proceed fairly. Indeed, if I treated the permissive rules of movement in chess as absolute truths and tried to combine them with physical facts, sometimes they would result in contradictions. (For example, the rules of chess would tell me that this rook can be moved forward one square, but the rules of physics would tell me that it cannot, because it is glued to the board.)

8.7. A General Principle of Cognitive Architecture

We are now in a position to tie together our observations about neural localization and domain-specific reasoning with the discussion of the disunity of science in chapter 7. These can be unified by hypothesizing that *it is a general design principle of the cognitive architecture of humans (and other animals) that the mind possesses multiple models for understanding and interacting practically with*

different aspects of the world. Having such mental models is an important factor in allowing animals to move beyond the control of the stimulus by understanding the world through "offline" representations of types of situations in the world, and the possible events and actions that they afford. Something is never simply "a model of the world," however. To be a model, it must be a model that represents the world as being *thus*, where the "thus" stands in for the possibly complex and proprietary subjective ontology of the domain of the model, the stock of properties and relations it models, the "dynamic principles" by which one recognizes possible changes in the world, and the set of possible interventions the organism can make. That is, to have a model at all, it must be a model that at least implicitly involves a representational system with a particular formal "shape." (Of course, it need not be a symbolic or linguistic representation. The model may not be explicitly encoded as a set of rules, even in a "language of thought," but may be implicit in dispositions of thought and perception and the neural structures related to them.)

Moreover, to divide the world up into objects, properties, relations, events, and the like, a model must deal at an abstract level, picking out specific aspects of the world to be represented, and conceiving them in particular ways. As a result, each individual model will deal with only a pared-down set of the numerous complex factors that are at work in real-life situations. That is, the model will be *idealized*, and sometimes the idealizations will matter in the sense that predictions based on a model will sometimes go awry.

There are two major types of constraints on what types of models a mind can possess. One consists in *what the neural resources will allow*. A brain that is built in a particular way can do some things routinely, other things only with difficulty, and still other things it cannot do at all. The second set of constraints is *pragmatic*. A model is adaptive to the extent that it allows an animal to do things well enough for biological purposes. Even in human minds, whose interests range widely beyond biological goals of nutrition, safety, and reproduction, models still answer to practical constraints of how well they serve their specific tasks, even if those are tasks of theoretical understanding.

Given this basic design strategy for cognitive architecture, architectures answering to this strategy can be, and have been, implemented in a variety of ways. The oldest way is through the production of special-purpose neural circuits through genetic variation and selection. Such a process is slow but stable. It affords "out-of-the-box" (or "out-of-the-egg") functionality that does not require trial-and-error learning. A slightly more complicated strategy is found in brains that exhibit more plasticity in development and can fine-tune to maximize the utility of brain tissue in different circumstances. Such brains, however, tend to have strong biases toward particular developmental trajectories, and major departures from these are mainly compensations for damage or abnormal growth. A third strategy is Anderson's massive redeployment of existing brain regions in new configurations. A fourth, and a major turn in cognitive evolution, consists in the evolution of brains that are not limited to an innate or developmentally biased set of anatomically localized models. Animals with such brains can gain new concepts and develop new models "on the

fly.'' This affords such animals with far greater adaptive potential. But learned models tend to operate more slowly—sometimes too slowly for adaptive response. However, because evolution tends to be *conservative*, adding new functionality without discarding the old, more sophisticated organisms tend to have cognitive systems of both ''older'' and ''newer'' types: for example, hardwired species-constant circuits like that underlying the blink reflex as well as abilities to judge and even theorize about trajectories of moving objects. Some such learned models are learned through, and partially constituted by, social interaction, including supervised instruction through language. This, in turn, provides a basis for a special type of such models that is developed in technological communities and employs particularly rigorous explicit representations in natural and mathematical languages: the sciences.

All of these share characteristic features of models: they are partial, idealized, and pragmatically driven, and they employ proprietary representational systems. These factors alone—the features of models *as* models—are capable of creating barriers to integration of models, including irreducibility and inconsistency. Thus, the hypotheses that human thought is model-based, and that scientific thought is a special case of model-based thinking, provide candidate explanations for the disunities of both ordinary thought and the sciences. They also entail at least the possibility that such disunities are an unavoidable artifact of our cognitive architecture. I view it as an open question whether there are types of minds that could know as much as we do and yet accommodate it within a unified representational system. It could be that the idealized nature of models implies not only the *possibility* but the *necessity* of such disunities, either on *a priori* grounds or when faced with the task of modeling a world like ours. And *if* it is possible that *some* beings could have a more unified mental life, it is an empirical question whether *we* can do so. Possible barriers to this could stem from the peculiarities of human minds and brains (e.g., if they contain two or more innate systems that are incommensurable or inconsistent) or from facts about the relation between our particular representational capacities and the things we try to understand in the world (e.g., if our minds are well-suited to understanding only classical objects, and hence not fully capable of understanding the quantum world or reconciling it with the classical).

Of course, cognitive architecture is not the *only* possible source of disunity. It could well be that additional disunities are forced by other factors. Sometimes two models are at odds with one another only because one or both of them are not the *best* models we are capable of coming up with for the phenomena they attempt to model. Ptolemaic cosmology and Cartesian physics involve principles that are at odds with contemporary science at least in part because they got things *wrong*. Likewise, because our language, social customs, and scientific practices are constitutively *social*, there could be social factors that generate disunities as well. For example, different cultures might have hit on very different ways of understanding things like the relation between individual human beings and their families and communities, and these might lead to their having words and concepts that not only resist ready

translation, but also cannot be accommodated within a single model of the social world. All of this, however, is consistent with the thesis that cognitive architecture is *one* important source of disunity, and perhaps even of disunities that cannot be overcome.

8.8. The Plausibility of the Story

This general story about the principled disunity of cognition seems reasonable on a number of grounds. One way of supporting it is on grounds that are aprioristic or perhaps transcendental. To represent the world is to bring it under concepts.[6] Concepts are abstractions from the rich and noisy mix that is the real world. It is *good* to represent the world abstractly, because this is what makes learning and inference possible. It would be a very maladaptive organism indeed that could only observe particulars and not learn, say, from one encounter with a tiger that other tigers are to be avoided (or perhaps at least as usefully, that anything that looks at all like a tiger should be avoided, as it is better to avoid tabby cats than to be eaten by leopards). Organisms have limited information-processing capabilities, and so an efficient mind is one that screens out information that is not relevant to the organism and flags information that is likely to be relevant.

I like to make this point through a creation myth. Suppose that God, on the fifth day of Creation, delegates two demiurgic angels to create some herbivores. One creates deer. Deer are relatively stupid: they take virtually any moving object to be a potential threat, and take even the snapping of a twig as potentially indicating a predator. Their predator-detectors are thus coarse-grained and generate many false positives (i.e., representations that there is a predator present when in fact there is none). But this saves them from getting eaten by tigers often enough that they proliferate. The other angel creates unicorns. Unicorns are created much smarter than deer. They recognize a lot more about their environment and spend a lot of time thinking things through. When a twig snaps, they wonder whether it is really a predator, and whether all tigers are dangerous; they entertain Cartesian doubts about whether their tiger-images are actually caused by malicious demons, and so on. In the process, they stand around thinking while the tigers are leaping, and as a result they all get eaten by tigers, and so you have never seen one (as the song goes) to this very day. The moral of the story is that too much information, or too much information processing, can be maladaptive. An organism designed *primarily* for reasoning, or even for gathering as much information as possible, regardless of its salience, would likely end up extinct in a world like ours.

This leads to a second argument for this account, an argument from evolutionary biology. What evolutionary biology predicts is that organisms are not endowed with cognitive systems optimized for reflecting the world exactly as it is, in all its detail, but for pragmatic purposes.[7] Both hardwired special-purpose systems and mechanisms underlying learning are built out of

a limited stock of neural components and are selected because they are *good enough*. Moreover, evolution is conservative and aggregative: phylogenetically older systems are generally not discarded when newer systems are added. Rather, the newer systems either supplement the old (hence many cases of redundancy as newer and more fine-grained systems are added to older quick-and-dirty ones) or built to co-opt the older ones. The brains of sophisticated animals are thus characterized by a significant level of *redundancy* as well as *hierarchy*.

There is no doubt but that mind and brain took a radical leap in genus *homo* (probably not the only such leap, but in some ways the most dramatic). There is much debate over comparative roles played by various gains made in this transition: more processing power, new types of cells, greater numbers of cells, higher interconnectivity, language, tool use, and various aspects of sociality. However, it is doubtful that the changes that made humans so different from other animals took a form on the order of scrapping all vestiges of animal proto-thought and endowing the human mind with a stock of innate ideas that reflect the real essences of created things, as the Rationalists would have it. Indeed, even a believer in special creation by God on the sixth day would have to take note of the striking continuities between human and animal cognition: at the very least, God built people out of many of the same neural parts as He used for chimpanzees, deer, alligators, and frogs. The question, then, whether one likes evolutionary or theological accounts of the design of the human mind and brain, is whether the special abilities that separate human and animal cognition involve something completely different from the pragmatically driven, neurally constrained, perception-mediated, idealized models employed in the older forms of cognition, or something so radical as to allow the construction, through learning, experiment, and theory, of a well-unified mirror of nature that could ultimately give us all that the Rationalists wanted, albeit in a nonnativist form. Those theists who are inclined to believe in special creation, especially if they are dualists, may still have some reason to expect the former possibility.[8] For those who prefer evolutionary explanation (regardless of the rest of their theology or atheology), it seems reasonable to adopt the more conservative option—that human minds are in large measure doing the old, adaptive things in new and more powerful ways—as the default assumption.

A third reason for liking this cognitive pluralist story is that cognitive triangulation through diverse mental models—especially when augmented by the uncontested human ability to coin new models, and to revise and test them—has important advantages over a more unified and pure cognitive architecture employing a single representational system. Kant suggested, for example, that intuitions of space are innately constrained by a Euclidean representational system. Kant saw this as providing the grounds for synthetic *a priori* knowledge of the phenomenal world (which, for him, was the world studied by science as well as the world of everyday experience). But a being that was really like the beings Kant described would have an important limitation. Suppose that we were to build a model of a Kantian cognizer in

the form of a robot that represented its spatial surroundings in some particular way, say, as an orthogonal grid of a fixed resolution and extent (say, 3×3 or 1,000×1,000) with a particular spatial resolution of, say, one pixel to the meter. Just to make the example more stark and tractable, let us make it a two-dimensional grid. (The example is not so bizarre. Some animals, like the aforementioned deer, seem oblivious to threats significantly off the ground, suggesting that their representation of space may have dimensional constraints that ours does not, making them more vulnerable to predators hunting from trees, like jaguars, pythons, or human hunters in elevated blinds.)

This robot, which we may call Ortho, can represent things well enough so long as they are on the ground and do not require a spatial resolution more exact than one meter. Moreover, some sorts of computations, such as figuring out what is next to what, simply fall out of this sort of spatial model with little computational difficulty. For some problems, this representational system is actually *elegant*. It might be a *good* system if Ortho is primarily designed to play chess or tic-tac-toe on an appropriately sized grid. But there are other things that it cannot represent aptly, such as things whose elevation is significant, or anything requiring a greater degree of spatial resolution, or objects lying exactly across the boundaries between spaces in its grid. The trouble is that, by Kantian lights, it is a synthetic *a priori* truth for Ortho that all objects are located by reference to its particular sort of spatial grid. Yet from our perspective, what it gets in such cases is not truth at all, but *forced error*. For a truly Kantian cognizer, it is impossible to tell the difference between synthetic *a priori* truth and forced error (cf. Cummins 1996 for a discussion of forced error). This point depends in no way on the *particulars* of how Ortho represents space. It would hold true if its representational system were three-dimensional and had one hundred times the resolution, or if it used polar rather than orthogonal coordinates. The details of the forced errors would be different, but the philosophical point would remain.

But suppose that we were to endow the next generation of robot, Pluro, with *multiple* representational systems. We might, among other things, give it both orthogonal and polar systems for representing space. These might be apt for different tasks. (An orthogonal model is elegant for representing the chess board before it; the egocentric polar coordinate model for calculating how to reach out and move a piece.) Moreover, there might be ways that Pluro could combine the insights of different models to provide information that could not be extracted from any of them individually. This kind of informational triangulation would provide it with a level of sophistication unavailable to Ortho's single spatial representation system. But these would *not* be based in some representational system that reduced the two to a common denominator. Moreover, if we additionally gave Pluro the ability to coin new and more adequate representational systems on the fly, and perhaps to improve things like the granularity of its spatial information sensitivity through the use of instruments, it would gain yet additional advantages. Such an intelligent and pluralistic architecture has distinct advantages over a

more unified Kantian architecture, as it confers the ability to recognize and perhaps transcend individual sources of forced error inherent in any one representational system.

Moreover, there is good reason to suppose that we are a good deal more like Pluro than Ortho. We *can*, for example, pose and investigate the question of whether space is non-Euclidean, even if our innate perceptual endowment *does* represent things in a Euclidean geometry.[9] And we can recognize that the objects of quantum physics do not always behave the way our deeply ingrained categories of *substance* and *causation* would have them behave, and even build mathematical bridges to a tenuous grasp on such objects and their behavior, even if it is not grounded in a spatial intuition. So the pluralist story about cognitive architecture is not only advantageous, it would seem to correspond to the evidence as well.

8.9. Pluralistic Cognitive Architecture and the Sciences

Given the plausibility of such a pluralistic cognitive architecture as a general story about human and animal minds, what story should we tell about the sciences as cognitive enterprises? One possibility, of course, would be that these are radically different from nonscientific thought in a way that underwrites the possibility of unified science as conceived by, say, Carnap or Oppenheim and Putnam. This might involve either a story about a new "theory module" that is an innate endowment of *homo sapiens* (even if originally selected for different purposes than the scientific enterprise), or perhaps even one that evolved very recently, roughly coincident with the emergence of theoretical thinking in humans, or a story about how a system best described in the ancestral environment as a special-purpose module has since been co-opted through social and linguistic processes into something that turns out to have new resources for unified theory. Alternatively, it might be that even scientific models, like other domain-specific knowledge, are particularly refined instances of special-purpose representations, albeit ones that are formed through learning rather than innately endowed through hardware modularity, and hence less securely tied to the particulars of evolutionary history.

I think there are a variety of reasons to prefer the latter hypothesis. First, the possibility of a *very recent* emergence of a hardware module for integratable theoretical thought is a nonstarter in evolutionary theory. The ability to engage in scientific thought is spread across the entire human species, in populations that have been genetically isolated over much longer periods than the stretch of time that has passed since the emergence of theoretical thinking, much less modern science, within the species. If there is a hardware module that operates on such radically new principles, it must be much older, dating to the emergence of *homo sapiens* or even genus *homo*.

But the grounding of unified science in such a hardware module seems unlikely in any case. First, it is difficult to see how such a mechanism could have been selected for such a purpose if it has been *used* for such a purpose

only very recently. Second, everyone would be in agreement that uniting our various strands of knowledge is an *accomplishment*, and an accomplishment *hard-won*. It is not simply guaranteed by cognitive architecture. We may be built, as Kant suggested, with a *drive to unify* knowledge, but by this very token we are also built in such a way that our knowledge starts out *dis*unified. We do *not* start out representing the world in a unified system. At best, we must revise, massage, and order what we know to achieve greater unity. And so, *at most*, theoretical reasoning is built in such a way that it *presents no intrinsic barriers* to unifying knowledge. But the fact that there is a *drive* to unify leaves it an open question as to how far such unification can actually be accomplished. As a limiting case, we might recall Kant's cautions about the dialectical illusions of the Ideas of Reason.

And so, even if we suppose that the organic basis for theoretical knowledge employs representational resources unlike those of the parts of the brain that are more modular because of their hardware-level descriptions, we must additionally suppose that these must be pushed in a particular direction through particular cognitive, experimental, and social practices if theoretical integration is to be achieved. There is undoubtedly *something* right about such a story. Human adults, especially those educated in particular ways, *are* capable of cross-domain reasoning in ways that seem to be unavailable to other species and even to young children (Sperber 1994; Carruthers 2002; Spelke 2002). The question, however, is how comprehensive a form such cross-domain reasoning—and more important, *unification*—can take. This is in large measure an *empirical* question about human psychology. Moreover, the pressing question at hand is not cross-domain *reasoning*, but *unification* of domains. Cross-domain reasoning can be had merely through the ability to deploy and move between *separate* and even *incommensurable* models to triangulate the world. Indeed, so long as it is truly *cross-domain* reasoning, the domains are *not* yet unified.

There are both empirical and theoretical reasons to think that thought is abidingly pluralistic. On the empirical front, we are confronted with evidence that different knowledge domains, including the sciences, have not been reductively unified or even reduced to a common denominator, in spite of centuries of attempts by the best minds of modernity. This at least suggests that they may not be subject to unification. There are also theoretical reasons for suspicion about our capacities for successful unifications of knowledge. If knowing always involves modeling the world in particular representational systems, and these are always idealized, there is always the formal possibility that the representational system that proves apt for one sort of problem might be incommensurable or inconsistent with another representational system that is equally apt for another sort of problem. Two physical models may each be highly apt in particular contexts, but may not be integratable. Models that employ normative concepts may be necessary in the social domain but not be constructible out of nonnormative resources. We cannot decide on purely theoretical grounds whether our minds are necessarily such that they must employ models that cannot be integrated with one another theoretically.

But the empirical evidence of an abiding plurality of domain-specific models, including scientific models, seems to suggest that we *do* have such an architecture. Or, at the very least, that our existing ways of thinking about the world, including our multiple scientific models, are not subject to wholesale integration into a single supermodel or God's-eye view of the world.

The die-hard reductionist or unificationist might hold out for the possibility that a substantial *reconception* of all our scientific and nonscientific knowledge might result in a single, grand axiomatic system. But I find such a hope unwarranted on several grounds. First, it is not at all clear that a system that was integrated in this way would retain all of the apt insights of our current hodgepodge of models. Second, it is not at all clear that we are *capable* of such an epistemic revision. Human beings, across cultures and history, seem to have hit on very similar ways of thinking about the world, and this is at least suggestive that there are innate biases toward forming particular types of apt and adaptive models. Such biases plausibly constrain not only "folk" knowledge, but scientific knowledge as well.

In short, it seems a likely hypothesis that our *de facto* plurality of theories and models is a consequence of *deep* facts about cognitive architecture. There is good reason to think that such plurality is in fact intractable. There are models of cognition that predict this result. And these models are plausible on a variety of grounds, including evolutionary grounds.

8.10. Conclusion

Earlier, I posed four questions that ought to be answered by a post-reductionist philosophy. This chapter and its predecessor have addressed the first, that of accounting for scientific pluralism. According to the cognitive pluralist account, this is a consequence of general features of human cognitive architecture, which produces models that are pragmatically driven, idealized, and constrained by the neural networks through which they are realized. Three other questions, concerning metaphysics and the intuition that there is something special about the psychological gaps, have yet to be addressed. These are the topics of the next chapter.

This page intentionally left blank

9

Cognitive Pluralism and Modal Metaphysics

In chapter 7, I posed four questions that a post-reductionist philosophy ought to address.

1. What is the origin of scientific plurality and failures of reductive explanation?
2. Is the Negative EMC tenable, and indeed how would we go about deciding such a question?
3. Are intuitions to the effect that the natural phenomena encountered in biology and chemistry supervene upon basic physics motivated apart from discredited reductionist arguments?
4. Is the intuition that there is something different and special about the psychological gaps defensible, and if not, can it be explained away?

The cognitive pluralist answer to the first question was addressed in chapters 7 and 8. Three questions remain.

9.1. Cognitive Pluralism as Metaphysics

The previous chapters dealt with Cognitive Pluralism solely as a claim in philosophical psychology and epistemology. The epistemological side of Cognitive Pluralism is something that many physicalists and dualists might find congenial. Indeed, the appeals to brain architecture and evolutionary considerations are exactly the sort of thing that many physicalistic naturalists might well applaud. (It is, indeed, a ''naturalistic'' project in the sense in which that term is used in

epistemology and philosophy of science, which is importantly distinct from the usages in metaphysics.) Even dualists need not deny that the mind employs multiple models that represent features of the world in particular ways, nor maintain that these capacities are wholly independent of the architecture of the brain. Descartes, you will recall, postulated several mental faculties such as imagination and thinking, and took very seriously the interactions between mind and body as shaping much of what we would call psychology.[1]

Indeed, in one sense Cognitive Pluralism is not a thesis about metaphysics or ontology at all, and is ecumenical between rival views. That is, it is not a thesis about what I have called "*positive* ontology" or "inventory ontology": the question of what is included in the inventory of the world. In exactly the same way, the Kantian and Pragmatist models of cognition are not (positive) metaphysical or (inventory) ontological theses, and hence both materialists and dualists have on occasion availed themselves of Kantian and Pragmatist epistemologies or philosophical psychologies.

Yet in another sense Cognitive Pluralism can, like Kantianism and Pragmatism, be taken as a theory about metaphysics and ontology—not in the sense of positive or inventory ontology, but in the sense of "*critical* ontology" as "the study of being *qua* being." Kantianism and Pragmatism *do* have things to say about what it is to be an object: namely, that it is to stand (or potentially to stand) in a particular sort of relation to cognition, practices, or interests. Unlike Berkeleyan idealism (at least as understood by Kant and many others), which treats objects as *merely* perceptions and hence has startlingly revisionary implications for inventory ontology as well, part of the genius of Kant and the Pragmatists was to leave such questions essentially untouched. The inventory of the (phenomenal) world can consist of whatever our best scientific understanding (and perhaps some nonscientific understanding as well) says it consists of. What they are interested in is the status of objecthood.

Thus what they are rejecting is not something on the order of materialism or dualism, but a kind of deep *realism*, which their followers are sometimes wont to call "naïve realism." Naïve realism is the uncritical assumption that the world divides itself, in a unique, canonical, and mind-independent way, into objects and properties. Hence, for the naïve realist, the job of the mind is to adequately reflect these "real natures" or Lockean "real essences." One may, however, endorse this realist view critically (reflectively and nonnaïvely) as well, and so I shall use the term 'realism' here stipulatively, as indicating the view that the world divides itself in a unique, canonical, and mind-independent way into objects and properties. Realists, as it were, accept part of Quine's (1953) dictum that "to be is to be the value of a bound variable" (albeit perhaps only for bound variables in the theories of an ideally completed science), without pondering the implication that this definition of objecthood—being the value of a bound variable in a particular type of theory or model—itself cashes out objecthood in terms of models (and hence the psychological activity of model*ing*) that involve things like variables and quantification. The naïve realist is simply unaware of such a further issue. The more critical realist is aware of the question, but takes the posits of such theories as ontological bedrock. The

posits of such theories are as deep as we can go, says he, and hence there is nothing to ontology except positive ontology.

Kantians and Pragmatists reject the realist assumption that the division of the world into objects is independent of minds, practices, interests, or conceptual schemes. Cognitive Pluralism shares this view, but additionally rejects the hypothesis that the division is unique or canonical. Quine (at least in some of his writings; Quine 1969) is thus in some sense a cognitive pluralist as well as writing in the Pragmatist tradition. But Quinean pluralism comes in the form of multiple frames of reference or conceptual schemes, and these are posited as completely *global* ways of carving up the world. Cognitive Pluralism of the sort I am proposing, by contrast, is skeptical of the supposition that there *is* anything as comprehensive and consistent as a Quinean conceptual scheme or frame of reference. Rather, there are a number of local models of portions or aspects of the world, each of which assumes its own positive ontology, and it is at best an open question whether these can be integrated into something both comprehensive and consistent.

There is a possibility that such a view could result in *some* implications for inventory ontology after all. For if the criterion for objecthood is being the postulate of an apt model, and aptness is dependent on context and of practical and explanatory interests, then it may well turn out that models employed in various everyday contexts will turn out to be just as respectable, by these lights, as those of fundamental physics. If this is the case, then the inventory ontology that results may be far more profligate than, say, its Quinean counterpart, allowing thoughts and even royal flushes to be as respectable as quarks and leptons. Whereas Quine's inventory is that of a desert, or rather an infinite array of parallel deserts corresponding to each conceptual scheme, such an inventory would be that of a rain forest. I am not sure that it is obligatory for the cognitive pluralist to move in this direction, however, and I shall not pursue the matter further here.

9.2. Modal Metaphysics

If Cognitive Pluralism is arguably ecumenical with regard to inventory ontology, it is not clear that it is so neutral with regard to modal metaphysics. Or perhaps better, it raises problems for certain *interpretations* of modal metaphysics that are standardly in use. In particular, it may raise problems for a modal realist version of possible-worlds semantics (PWS) generally, and for an integration of PWS with scientific theories in particular.

Here is a familiar, middle-of-the-road way of developing the notion of possible worlds. Take the set of all propositions. To each of these there corresponds a state of affairs: an abstract entity corresponding to a way the world might be, locally. Now define a function mapping the set of propositions onto truth values. Each such mapping picks out a global state of affairs, which we may designate a "world" (again understood abstractly as a way things might be, or might have been). Not all assignments of truth-values are consistent. If

we assign TRUE to "There is at least one dog" and FALSE to "There is at least one animal," we have a contradiction. Worlds corresponding to such an assignment of values have incompatible states of affairs and are not really possible. Not even God could (*pace* Cartesian voluntarism) bring such a world into being. So let us designate worlds as *possible* worlds just in case they consist in jointly compatible states of affairs and correspond to a consistent assignment of truth-values to the set of propositions. A proposition P is necessarily true if it is true at all possible worlds, necessarily false if it is false at all possible worlds, and possibly true if it is true at one or more worlds. A state of affairs S is necessary if it obtains in all possible worlds, possible if it obtains in one or more possible worlds, and impossible if there are no worlds in which it obtains. Particular modal logics (such as T, S4, and S5) are distinguished by axioms regarding necessity, possibility, and counterfactuals.

Metaphysical supervenience is standardly explicated in possible-worlds semantics. A state of affairs S1 metaphysically supervenes upon S2 if S1 obtains in all worlds in which S2 obtains. (Sometimes some additional nonlogical dependence is also thought to be required, else all necessary states of affairs would supervene upon all other states of affairs.) A property P1 supervenes metaphysically upon P2 if it is the case that all worlds in which x is P2 are also worlds in which x is P1. (More sophisticated formulations are required if P1 and P2 are applied to different objects.)

Can a cognitive pluralist embrace modal metaphysics and supervenience, thus cashed out? It is a tricky question. At one level, the cognitive pluralist is bound to treat modal metaphysics and possible-worlds semantics like any other models. If there are problems to which they are aptly suited, more power to them. But at another level, PWS is problematic if it is taken as revealing the deep, fundamental, and mind-independent structure of metaphysical reality and as a canonical tool for revealing metaphysical truths. It is problematic both from the cognitivist perspective and from the pluralist perspective.

Cognitivist suspicions are aroused by the too easy use of expressions such as "the set of all propositions" or "the set of all states of affairs." This smacks of the very type of realism that the cognitivist, *qua* cognitivist, rejects. If this machinery requires us to assume that the world divides itself in a canonical and mind-independent way into states of affairs, the cognitivist ought to reject this. But perhaps this problem is not insuperable. Perhaps the talk of "all propositions" can be cashed out in terms acceptable to the cognitivist: say, as "the contents of all possible judgments (for minds like ours)," and likewise "all states of affairs" as "the states of affairs corresponding to all possible judgments (for minds like ours)." This would render PWS acceptable to the cognitivist. However, it is not clear that it gives the modal metaphysician all that she is looking for. If we restrict the space of propositions and states of affairs to things human minds are capable of conceiving, we are in danger of leaving out things that our minds are *not* suited to conceiving. And unless one takes the (unlikely) view that there is nothing that is beyond our ken, this falls short of what one needs for genuine metaphysical necessity and possibility.

(One cannot help but hear Kant's ghost complaining of attempts to speak of propositions or states of affairs outside of the realm of possible experience, and the conflation of phenomena with noumena, even if advocates of PWS do not take themselves to be talking about noumena, or even reject that Kantian machinery altogether.)

Qua *pluralist*, the cognitive pluralist finds PWS problematic on other grounds. First, the notion of *truth* in a cognitive pluralist philosophy needs further explication. The cognitive pluralist needs to distinguish between two levels of epistemic/alethetic goodness. One level is the *aptness* of a given *model*. It is only when one is dealing with a particular apt model as a background assumption that the second-level question of the truth of a proposition can be raised. Imagine that someone says, "The car was standing still on the tracks and the train came speeding into it." As courtroom testimony, this might rightly be regarded as a true statement. And it would do no good for opposing counsel to object that the statement is clearly false because there is no such thing as absolute rest. But in another context, where the question of absolute rest was the issue at hand, such an objection would be pertinent.[2]

This problem may not be insuperable, however, if it is possible to differentiate *statements* from *propositions*. One might try to make the simple assignment of truth-values to propositions compatible with pluralism by treating propositions as a function of [statement, model] or [statement, model, context]. On such a view, what is different in the two cases is first and foremost the proposition expressed: the *statement* "The car was standing still" would be taken to express a different proposition depending on whether it was a statement made in traffic court or in cosmology.

More troubling, it seems to me, is the notion that scientific claims (and perhaps all claims) are ultimately *idealized* claims. For it is not clear that idealized claims are sufficient to the task of underwriting either necessity claims or determinate values for counterfactuals. The problem is nothing unique to mind-body relations, for it is not clear that even the real-world kinematics of physical objects supervenes upon the combination of laws and prior positions, as Laplace would have had us believe. I shall not address potential challenges to the Laplacean vision that derive from quantum indeterminacy or classical chaos, but rather will concentrate on the kinds of issues raised by Nancy Cartwright (1983) about the integration of separate dynamic laws into descriptions or predictions of real-world kinematics.

Scientific models, taken individually, do not license exact predictions of real-world kinematics, because each one leaves out the contributions of the others. This, however, might not seem an insuperable problem, because it should be possible, in principle, to factor a real-world problem into components that each separate model (e.g., of gravity and electromagnetism) can describe, and then do a summation of forces through vector algebra. There might still be problems in chaotic cases in that our *calculations* must be done at a finite level of approximation, but this is no challenge to factoring and reintegration *in theory*. It is the sort of thing that a Laplacean demon could overcome if it was not confined to finite approximations of the values of

constants. (Though it is admittedly hard to see what its calculations would be like, if not represented by decimal sequences. See Section 7.5.1.1.) So long as one is dealing with independent forces, factoring is innocent at a theoretical level.

What *is* a problem for factoring, even at a theoretical level, are the other sorts of idealizations that are often at work in the background of theories: for example, the simplifying (and distorting) assumptions that bodies are point-masses or that collisions are perfectly elastic. Even worse are cases (e.g., of a portion of a feedback system) where a model treats things as independent when they really are not. *This* kind of idealization creates principled problems for reintegration of what has once been factored. In these cases, predictions and descriptions based on laws and prior conditions will fail to reproduce real-world kinematics.

Philosophers of science like Cartwright have addressed this issue at the level of scientific theory and prediction. But it can also be cast at the level of metaphysics. If by "scientific laws" we mean the sorts of idealized things we encounter in real sciences, then real-world kinematics does not supervene upon the combination of laws and prior states, even in the case of physics. It seems to me that there are two ways one could go with this. One course would be to reprise Locke's move to accommodate "real essences" and hold that scientific laws as we have them are not the *real* laws upon which (in conjunction with prior states) kinematic facts supervene. The *real* laws might be unidealized laws, tantamount to Lockean real essences. However, if idealization is a general feature of our cognitive modeling, we must then also join Locke in holding that real essences are unknowable to us, even if they would be knowable to God or Laplace's Demon. The alternative is to hold that the kind of laws we have are real laws, but that kinematic facts do not supervene upon them. This ensures whatever virtues accrue to epistemic naturalism, but at the cost of driving a wedge between science and modal metaphysics. One might have thought, after all, that (regardless of what one thought to be the best interpretation of laws themselves) once one had fixed the laws and prior states, the consequences thereafter were metaphysically necessary: that is,

$$L(A \rightarrow B) =_{df} A \rightarrow B \text{ in all worlds in which L is a true law.}$$

But if laws are idealized, specifying L and A underdetermines the truth value of $A \rightarrow B$ at a given world, and hence this characterization of laws (or "nomic necessity") will not do.

What should we make of this? I do not think we should conclude either (a) that there is anything wrong with our laws, or (b) that there is anything incoherent about modal metaphysics. But I do think it means there is a problem about trying to graft one onto the other. This does *not* imply that there is a problem with either of them individually, so long as one takes the cognitive pluralist line that scientific models and modal metaphysics are separate cognitive-pragmatic enterprises. Each can be apt for certain problems without it following that they can be united into a single coherent system. The

pluralism that was posited at the level of individual scientific theories can be put to work at the level of relating scientific theory to metaphysical enterprises as well.

9.3. Negative EMC

All of this would seem to have implications for how we ought to think about the plausibility of the Negative Epistemology-to-Metaphysics Connection Principle (Negative EMC) traditionally employed by dualists to argue their point, and also about intuitions concerning supervenience generally. Negative EMC is the thesis that principled failures of reducibility imply failures of metaphysical supervenience as well. I have always found Negative EMC to have a great deal of intuitive plausibility. Of course, to apply such a principle to a particular problem, such as mind-brain supervenience, one needs to project well beyond the current state of scientific knowledge. Negative EMC is not a claim about what follows simply if we do not happen to know how to reduce mind to brain today or tomorrow. It is a claim about what follows if such reductions are unavailable in a principled and abiding way. And there is always some risk in assuming that today's lack of understanding is principled and abiding. Those worries, however, are not about Negative EMC itself, but the other premise needed to yield an argument against physicalism: that is, that the mind is not simply unreduc*ed* but unreduc*ible*.

Negative EMC itself seems plausible on the basis of a very old and familiar way of looking at logical or metaphysical necessity: namely, that the denial of a necessary truth results in a contradiction. Thus one august method of proving necessity claims in mathematics and philosophy is the Principle of Noncontradiction. To prove P is necessary, assume not-P and show that it results in a contradiction. This methodology indeed informed much of the very conception of metaphysical necessity in early modernity. Hence there is a strong philosophical intuition to the effect that, if A→B is necessarily true, A-and-not-B should demonstrably result in a contradiction. In the case of psychophysical supervenience, if physical state P entails conscious state C, P-and-not-C should either be itself contradictory or generate a contradiction when combined with other necessary truths. To the extent that there seems to be no *inconsistency* involved in imagining Chalmers's zombies (as not possessing qualia or consciousness) or Searle's Chinese room (as not involving understanding), there is reason to think that there is no necessary connection between brain states or functional states on the one hand and consciousness or intentionality on the other.

There are several potential problems, however, for the *prima facie* plausibility of Negative EMC. The first, again, is that what we are talking about here is whether contradictions can be derived given a *complete* knowledge of the objects and properties under consideration, plus a knowledge of any other relevant necessary truths. Because no one thinks we are in such a situation, our intuitions on such matters are by no means completely secure. Unless one

knows a lot about microphysics, one does *not* see the inconsistency of imagining a one-ton block of plutonium. Unless one knows a good deal of classical thermodynamics, one does *not* see the inconsistency of imagining a change in mean kinetic energy of gas molecules without a change in temperature.

The exception to this rule is when (a) we are dealing with concepts that are pretty well nailed down and (b) one domain does not possess the *type of conceptual resources* needed to derive the concepts of the other. When I say the concepts are "nailed down," I am thinking of this: paradigm "natural kind" concepts like WHALE or WATER are in some ways open-ended. This is, in part, because they contain a large *ostensive* element. Their sense is, roughly, "*that* kind of stuff, whatever it is, that I recognize through the surface properties of being [description]." Such concepts are intrinsically open to unimagined ways of filling in the underlying nature of the stuff thus picked out, and the nature of what they refer to is underdetermined by their descriptive content. Other concepts, such as CIRCLE, behave quite differently. One does not need to investigate nature to know the nature of circles; it can be discovered (or perhaps stipulated) independently of experimentation. Likewise, if the foregoing gloss on the sense of natural kind terms is correct, there must be some preexisting stratum of more descriptive concepts that are not picked out in the same way, like CLEAR or SOUR TASTING or MOVING. These may ultimately admit of some revision in light of empirical experiments, thought experiments, and linguistic analysis, but of a much more limited kind than WATER or WHALE. They are already better nailed down than WATER, though perhaps not so much as CIRCLE. A number of writers, such as Kripke, have taken the view that the concepts that pick out phenomenal properties, such as PAIN, are well nailed down.

What about the lack of even *candidate* explanations? Consider the dictum that you cannot derive *ought* from *is*. This might be plausibly seen as a claim that there are fundamentally normative notions that cannot be constructed out of nonnormative notions. Or, alternatively, the claim that object-talk cannot be a construction out of sense-datum talk. Or in mathematics, that some particular system cannot be deemed a conservative extension of another. Proponents of the explanatory gap also tend characteristically to hold that there is something about *experience*, *meaning*, and *normativity* that resists being reconstructed in terms of the conceptual resources used to describe structural and functional properties—that this is like trying to view sound as a construction out of color.

As far as these considerations go, I am inclined to think that the weight of argument favors Negative EMC and its applicability to the mind-brain relation. Indeed, I am inclined to think this about intentionality as well as consciousness, as I think intentionality intrinsically involves at least the possibility of consciousness (compare Horst 1996; Searle 1992; Siewert 1998; Horgan and Tienson 2002).

But there is also a deeper concern here that is raised by our previous discussion of Cognitive Pluralism and necessitarian metaphysics. Most fundamentally, this metaphysical project is generally cast against realist and unificationist

assumptions that Cognitive Pluralism rejects or is at least skeptical of. Negative EMC is plausible insofar as one assumes that reasoning based on our concepts is a good way to investigate the real and fundamental natures of things in themselves, and hence to uncover deep metaphysical truths. In short, it relies on a good measure of realism and Rationalism, albeit not of nativism. Cognitivism, on the other hand, treats our concepts, not as reflecting Lockean real essences or Kantian noumena, but as things-as-represented-in-partial-and-idealized-models. This considerably deepens the suspicion of our intuitions based on conceptual analysis. *Now* the issue is not simply whether our *present* concepts reflect an adequate understanding of their objects, but whether *any possible* concept could do so. A failure to derive a contradiction might be an artifact of the representational systems employed. Conversely, the generation of a contradiction might be an artifact of a mismatch between two representational systems that cannot be smoothly integrated with one another. Indeed, the appearance of a reduction might even be an artifact of the ways representational systems have been idealized. To the extent that necessitarian metaphysics is to be about things-in-themselves and not things-as-represented-in-model-M, a cognitivist view of our concepts ought to engender suspicion about just how far exercises in necessitarian metaphysics can take us.

But perhaps we should not assume that necessity claims are supposed to relate unknowable real essences or things-in-themselves. Perhaps they are supposed to be limited to what Kant called the phenomenal world. Indeed, at least *some* of these Kant took to be *entailed* by cognitivism: the pure Forms of Sensibility and Categories of Reason. But this, too, seems problematic in at least two respects. As we discussed earlier, what looks from the inside like synthetic a priori truth might sometimes be reasonably regarded as forced error. The way we moved beyond limitations of particular forced errors was to posit multiple representational systems that could be used to compensate for, or even correct, one another. But this pluralism presents further problems for necessitarian metaphysics. In chapters 7 and 8, I made a case that separate representational systems can produce inconsistencies, not as a result of anything about the world, but as a result of the representational systems themselves. In this respect, there are limitations to what one can infer from the presence or absence of a contradiction. Moreover, if conceptually based necessity is really simply what is necessary *within a particular model*, and models may differ in their internal structure, such necessities look less and less like things that are about the world—even the Kantian phenomenal world—and more and more like artifacts of the representational systems involved in each particular model. To take a Kantian example: Kant suggests that when we are doing theoretical reasoning, we are constrained to view every event as caused by other physical events, and indeed to assume a type of determinism. Let us assume he is correct in this, though I happen to think that his reasoning here is invalid and probably depends on a deep misunderstanding of the status of laws in Newtonian mechanics.[3] On the other hand, he claims that, when we are doing practical and moral reasoning, we are constrained to treat humans as agents at least capable of free action, unconstrained by any external necessities. Kant's

clever and perhaps consistent solution to this is to suggest that we are phenomenally determined but noumenally free. When thinking about causation, we are thinking about the phenomenal world; when thinking ethically, we are thinking of the self as a transcendental subject, and as noumenally free.

Part of Kant's analysis I endorse: scientific and moral reasoning require us to employ different models of the world, operating upon different principles. In dealing with scientific models, we are looking at the world through lenses that factor out normativity and anomic events and highlight causal invariants. I do not happen to think that scientific reasoning need exclude anomic causation; genuine randomness is a part of some scientific models. But more important, I think that the *plurality* of scientific models drives a wedge between a commitment to causation and causal invariants on the one hand, and to determinism on the other. To hold to the truth of the gravitation law is *not* to hold that objects will actually behave as a description invoking only what that law would predict, because there are other laws at work as well, and perhaps also anomic causes. The gravitation law is a partial and idealized claim, and a claim, not about how objects actually behave, but about one regular contribution to their behavior. Each law, taken alone, is absolutely agnostic about whatever other causal factors may exist. Indeed, even a commitment to a particular set of laws as the complete set of *laws* leaves open the question of whether there are other types of causation as well, such as quantum randomness or voluntary spontaneity (Horst 2004). Thus a commitment to the truth of scientific laws does not entail a commitment to determinism or the causal closure of physics. On the other hand, to reason morally is to treat humans as having the freedom to choose their actions. This involves deploying another sort of model (or perhaps a family of models) very different from those employed in the sciences. But these models, too, are idealized, and are not antithetical to holding that there may be other factors (say, neurochemical imbalances, neuroses, or lack of good moral upbringing) that might stand in the way of genuine freedom in any particular instance. "Ought" implies "can" only within the idealization class of the deontic moral model. It is not an absolute truth about human beings, because there can be interfering factors that the deontic model has idealized away from.[4]

For both Kant and the Cognitive Pluralist, it is a philosophical cardinal sin to reason from a truth-within-a-representational-system (phenomenal world) to a truth-full-stop (noumenal world). But for Kant this takes just one characteristic form: to mix the phenomenal and the noumenal by applying the categories to noumena or treating phenomenal truths as noumenal truths. For the Cognitive Pluralist, there is *not* one single phenomenal model of the world, but a variety of local models. Likewise, whereas Kant treats ethical reasoning as reasoning about noumenal freedom, the Cognitive Pluralist takes it as reasoning within one *or more* models of human action with normative elements. The cardinal sin is far more general: it is to treat truth-as-represented-in-a-model as truth-full-stop. It is possible that Kant would have been sympathetic to this view, as there is evidence in the Third Critique that he views biology as irreducible to physics and employing separate principles, and

even in the First Critique he warns that our notion of "the world" is a dialectical illusion. Be that as it may, the point is that reasoning based on partial and idealized models is problematic for projecting categorical claims about reality, and hence is reason to view Negative EMC with some degree of suspicion.

It is, by the way, equally grounds for viewing *Positive* EMC with some suspicion as well. Positive EMC is the apparently more innocent claim that a conceptually adequate explanation of A in terms of B entails that A is metaphysically supervenient upon B as well. But if A and B are, or are cast in terms of, partial and idealized models, this inference holds good only insofar as there is nothing about the idealizations in question that could undercut the inference. If, for example, B is a claim about gravitational *forces* and A a claim about real-world *kinematics* that represents the world as it would be if only gravitational force were at work, B is derivable from A, but can yield false predictions about the real-world behavior of objects even if A is true, because a kinematic model needs to respect *all* of the causal contributions.

9.4. Supervenience

Discussions of supervenience are in fact more generally problematized by Cognitive Pluralist commitments. First, we have reason to be suspicious of our *intuitions* about supervenience. *Why*, for example, do we so naturally assume that chemical facts supervene upon physical facts? This may turn out to be a consequence of some deep bias in models we innately, or simply as a matter of fact, apply in our reasoning about the world. But such biases may reflect (forced or unforced) errors or results of dialectical illusion. Consider, for example, our deep-seated intuition that the world consists of classical objects. This can seem like something that has the force of a truth of reason (e.g., that God doesn't play dice with the universe). But what is a truth of reason except a way we are constrained to think by our cognitive architecture or the models we happen to employ? The discovery that the most fundamental entities of contemporary physics do *not* behave like classical objects—and hence are very difficult to understand if not abidingly mysterious to minds like our own—should give us pause here. It may indeed reveal features of the deep architecture of human cognition—an empirical fact about our psychology—and the limits of its suitability to the task of understanding certain types of problems. Taking such intuitions, and those about supervenience, as reflecting deep truths about the world is a very risky undertaking.

Supervenience is also made problematic in a second way. Let us assume for purposes of argument that supervenience claims are clear enough if we assume a realist attitude toward metaphysics. But what do they *mean* if we reject this attitude in favor of a cognitivist, and particularly a Cognitive Pluralist, metaphysics? If they are claims about necessity-within-a-model, they are perhaps innocent, but they no longer amount to necessity-full-stop, and hence fall short of what is required for necessitarian metaphysics. But if they are something other than this, if they are on the order of claims about something

like Kantian noumena or Lockean real essences, it is not clear how we are to make sense of this. Supervenience claims are characteristically claims about necessary relations between *properties*. And properties are generally seen as the sorts of things that can be reflected by, or are the projections of, concepts. But noumenal supervenience claims cannot be grounded or reflected in concepts in this way. Such claims are either a form of Kantian dialectical error outright, or else require our understanding "properties" as something like unknowable (and indeed unthinkable) Lockean real essences. If this is the case, there may *be* supervenience relations, and perhaps God can know them, but we cannot know or even state them, in which case necessitarian metaphysics is an enterprise best left to God and the angels, and perhaps to any humans who have ascended to Platonic *noetic* insight. To the best of my knowledge, most philosophers do not fall into any of these categories. If we were to construe the problem in this way, we would be well served to take to heart Aquinas's reported assertion, after receiving the Beatific Vision, that he now regarded all his philosophy as filthy straw, or Plato's parable of the cave.

Let us then return to the question of what to make of our strong intuitions that, say, chemical facts supervene upon physical facts. On the one hand, these might be taken as a claim about physical and chemical *models*: that the facts stated in chemical models are derivable from those stated in appropriate physical models. This claim has two problems. First, it appears to be a false claim, as chemistry and other special sciences all seem to deal in facts that are *not* reducible to physical facts. Second, even if such reductive derivations *were* available, they would amount to metaphysical necessities full-stop only if there were nothing in the idealizations involved in such models that might render the derivation a mere artifact of the models—for example, if our chemical models screened out real chemical phenomena that were not derivable from the physical model.

On the other hand, the supervenience claim might be a claim, not about the phenomena-as-modeled, but about real and fundamental essences that might not be entirely captured by the model (thus perhaps accounting for the explanatory gaps between models as artifacts of how the properties are represented in the models). But to say this is to say something not very clear. It is, in fact, to point to things that cannot be *said* or even *conceived* at all. Nor, indeed, can they be tested. They cannot be tested empirically, because it is only models that can be tested. And they cannot be tested by analysis either, as analysis is concerned with what is implicit in the structure of a model. Realism can be saved here only by resorting to a Lockean move to unknowable real essences, and all its attendant skepticism.

Does this mean that necessitarian metaphysics is totally bankrupt? Not necessarily. For the Cognitive Pluralist, a model of some aspect of the world can be (and probably is necessarily) a partial and idealized story. Insofar as necessitarian metaphysics is viewed as another such model, or perhaps a project employing a number of such models, it may have apt applications, in which its theses have as much claim to truth as claims in the sciences or in everyday speech. One of the important idealizations implicit in such a project

is that it is only as good as the other models it works with. And insofar as it works across models, it is only good so long as it respects the idealization classes of those models. And it is, moreover, parasitic upon the way our understanding of the world is already encoded in other, more substantive models, such as those employed in the sciences. Moreover, it would seem that it is practicable only through the familiar techniques of derivation of truths within a model or reduction of one model to another. When it is practicable at all, both Positive and Negative EMC would seem to be appropriate postulates of metaphysical reasoning. However, reasoning based on those principles is trustworthy only to the extent that the models are apt and do not have conflicting idealization classes. And as we are characteristically very bad at evaluating such matters, metaphysical reasoning is risky business.

9.5. Why Is This Gap Different from All the Other Gaps?

Finally, let us consider the remaining problem posed in earlier chapters: the intuition, felt by many, that there is an important difference between the psychological gaps and the gaps one finds between the natural sciences. Does Cognitive Pluralism have any resources that will allow us to explain this intuition, or alternatively, to explain it away? My belief is that it does. Indeed, this issue has long been addressed by cognitivist philosophers such as Kant (1968) and Husserl (1913/1931), who have suggested that there is a special type of distortion involved in modeling the self and its experiences that is not found in modeling the objects of experience.

Recall that, for Kant, the soul, like the world, is one of the illusions of dialectical reason. This illusion results from treating the *subject* of experience, what Kant calls the transcendental ego, as though it were another phenomenal *object*. To treat something as an object, for Kant, is to apply the category of Substance to it. But the transcendental ego does not appear in experience as an object of thought—indeed, it does not *appear* in experience at all, but is the grounding of the transcendental unity of apperception (the ability to unite all my experiences as *mine*). Thought—or more specifically, judgment—for Kant is my application of the categories and empirical concepts to the manifold of sensations to constitute a (phenomenal) object. Subject*hood* is *not* involved in judgment as an *object* of thought. When we speak of "the self" or "the transcendental ego" in this context, we are not picking out a phenomenal object at all, and the experiencing self is in some sense logically prior to and distinct from the constitution of any object.

Husserl develops the point slightly differently. Thought, for Husserl, has the internal structure *ego-cogito-cogitatum*, or I/think/intentional object. Like Sellars (1960), Husserl does not view this structure as a *relation* between objects (the I, the thought, the intentional object). And like Kant, he does not regard the "ego" and the "thinking" revealed in transcendental phenomenology as objects. For him, "object" is whatever fills the object-slot of an episode of thinking, and is thus to be distinguished from what occupies the other slots. To

be sure, we *can* think and talk about people as "empirical subjects"—that is, as things to which psychological predicates apply. Likewise, we can "thematize" thoughts—that is, treat them as objects. But to do so is necessarily to distort their role in lived experience.

There is, I think, a common theme here that is endemic to the cognitivist/transcendental idealist view of subject and object. On this view, object*hood* needs to be cashed out in terms of *being a possible object of cognition*—and thus in light of the *I/think/intentional object* schema. Self*hood* or transcendental subjectivity is cashed out in terms of a different slot of this tripartite schema, and intentionality in terms of the entire schema. The schema thus enjoys a certain logical or transcendental priority over objects and even object*hood*.

If one adopts this philosophical stance, there *is* indeed a *special* problem in trying to account for subjectivity or intentionality in objective terms, a problem that does *not* arise in trying to account for one sort of object in terms of another. It arises from attempting to explain the first two slots of the schema in terms of particular objects that might occupy the third, and indeed to treat the elements of the first two slots *as* things that go into the third slot, the object-slot. Insofar as there is any bedrock to such an approach, it does not lie in objects at all, but in the structure of cognition through which objects are constituted. And so the whole project of trying to explain this structure by appeal to objects is fundamentally misguided.

This is at least *one* sort of barrier to explaining subjective experience (consciousness) and intentionality in terms of any sorts of objects whatsoever, be they physical, spiritual, or Platonic. The tripartite structure of experience is the framework in which objectivity (in the sense of object*hood*) arises, and hence is prior to it, and not to be explained in terms of objects and their relations. This does indeed explain at least a part of the intuition that there is something *unique* about the psychological gaps. However, it leads, not to dualism, but to something like transcendental idealism.

The Cognitive Pluralist is free, and perhaps even obliged, to accept much of this story. But as a *pluralist*, she may be hesitant about one particular aspect of it: the inclination to treat this analysis as a new form of foundationalism. She may well be inclined to question whether the transcendental story told by Kant and Husserl provides a new bedrock on which all other knowledge is to be grounded, preferring instead to view it as one apt story among others that might prove mutually informative, while giving ultimate pride of place to none.

In particular, she might be inclined to insist on the legitimacy of questions that transcendental idealists have traditionally considered themselves debarred from even posing. For example, in the Kantian form of idealism, one is prohibited from posing questions about the aptness of the categories themselves. One's *own* forced errors cannot even appear on the Kantian agenda. But surely it is sensible to pose such questions. And in a pluralist framework, it might be possible even to give partial answers to at least some of them. We *can*, for example, become aware of our propensity to fall prey to various illusions of perception, conception, and reasoning. And at least some of these can indeed

be explained, as empirical facts about human cognitive architecture, by appeals to facts about the human brain and its perceptual organs. Indeed, at least *some* of the features needed for transcendental subjectivity might prove explainable, however partially, in terms of systems in the brain that have the right formal structures to realize them. There are clearly types of cognitive architecture that do *not* have this structure—for example, systems that do not divide the world into objects. Perhaps there is a positive project that might bear fruit here as well.

The Cognitive Pluralist is likewise empowered to adopt a realist stance toward cognition itself when faced with problems that require it. In particular, it is necessary to think as a realist whenever one is posing problems of epistemic *fidelity*: of assessing whether a particular representational system R1 is apt for modeling some part of the world. To do this, one must treat that representational system as an object, and must moreover treat some other way of representing the world R2 *as canonical*, in order to assess what information is lost in squeezing it into R1. To be sure, it is impossible to do so *globally* in one's own case, as one cannot get entirely outside of one's own representational resources. It is easy, however, to do so with minds (or their robotic simulations) simpler than one's own, as shown by the exercise in Kantian robotology in chapter 8. But it is arguably also possible to do so with isolated components of one's own cognitive apparatus. We *can*, for example, explain why we cannot see certain portions of the electromagnetic spectrum, why we cannot distinguish between certain combinations of visible light called *metamers*, and why we are subject to particular visual illusions: namely, because our vision is realized through a particular sort of neural architecture in which information is lost or misinterpreted in completely understandable ways. We can do this because we have more than one mode of cognitive access to light, and not merely those innately endowed in our visual system, and can use the former to reveal the inadequacies of the latter. In so doing, of course, we do not thereby explain transcendental subjectivity or qualitative character. But we may, for example, explain why the human color space has the formal shape it has and why there is a phenomenologically pure yellow but not a phenomenologically pure orange (Horst 2005).

9.6. The Forest and the Trees: Metaphorical Variations

Cognitive Pluralism thus combines some of the virtues of transcendental idealism with an ability to give scientific explanations of isolated features of cognition, taken piecemeal. Different models of self and world can be mutually informative without there being any model that is ultimate bedrock (be it scientific or transcendental) in terms of which all the others are explained. Whereas many philosophers have followed Descartes in viewing knowledge as a tree in which the trunk springs from the roots, the branches from the trunk, and the leaves and fruit from the branches, it might be better to view knowledge as being more like a banyan forest. Banyan trees, unlike oak trees, are not

sharply demarcated from one another. One tree's branch may embed itself in the soil and become a new trunk without being separated from existing trunks. A banyan forest is thus not a collection of individual trees, but a complex network of organically connected trunks and branches. Likewise, various regions of knowledge—various local models—may well be richly interconnected, as opposed to being either separate and isolated or branches of a common trunk.

Even this metaphor, however, has its limitations. The interconnected banyan forest is ultimately causally derivative from a single trunk; but domains of knowledge may well start out separate and connect only later, like graftings of one plant onto another. And just as grafted plants change one another at the level of DNA, the interaction between different areas of knowledge may result in epistemic hybridization along the way. But graftings involve taking a sprig of one plant and grafting it onto the stem of another, so again the metaphor is not adequate. Imagine a grafting technique that grafted entire living plants, springing from separate seeds and trunks, onto one another, so that each was subtly transformed in the process, resulting in an interconnected forest that had many trunks that once were separate plants, now transformed into a new and different organism in which it is no longer possible to completely distinguish one plant from another because of the connections and mutual influences of one upon another, and in which no single trunk is the *Ursprung* of the entire forest. Here we perhaps have a better metaphor for a pluralist conception of knowledge.

The moral of this metaphor is that a *pluralist* conception of knowledge need not be tied to the simplistic notion that different models are completely isolated from one another. Such an idea has been explored, in a nascent form, within philosophy of science in the form of Darden and Maull's (1977) interfield theories. It surely admits of many further adumbrations. This calls for an expansion of the work of naturalistic philosophy of science into other domains of knowledge: examining the ways that they mutually inform one another, hybridize, and generate new models of the world. How such connections actually take place will provide further clues to the empirical facts about our cognitive architecture. As I hope this book has shown, examinations of cognitive architecture can provide important and even revolutionary challenges and resources to philosophy of mind.

10

Cognitive Pluralism and Naturalism

I wish to conclude this book by returning to the beginning: to a discussion of naturalism in philosophy of mind. I suspect that there is a great deal in my Cognitive Pluralist account of epistemology and philosophy of science that many self-styled "naturalists" will find quite congenial and appealing. Indeed, the discussions of the role of the mind in explaining scientific disunity combines elements continuous with cognitivist philosophy of science (e.g., the work of Giere 1988 and Nersessian 1992) with a pluralism that finds voice in writers like Dupré. And the general strategy of discussing epistemological issues in a fashion informed by the sciences of cognition seems squarely in the same camp with the work of epistemologists like Goldman (1986, 1992). All of these writers might reasonably be called "naturalists" in the senses operative in their respective fields. Likewise, some philosophers of mind who style themselves "naturalists" might see my approach as far closer to their own than those of dualists like Chalmers or even materialists like Kim, whose reflections are not in close dialogue with the sciences of cognition. Indeed, after hearing a talk in which I presented some Cognitive Pluralist ideas a few years back (albeit without reference to naturalism), Ruth Millikan has subsequently described my approach as being in the same "naturalistic" camp as her own, I assume in large measure because it attempts to apply the cognitive and biological sciences in the task of solving philosophical problems about the mind. I, on the other hand, have always viewed the thrust of this project as being a fundamental *critique* of naturalism in philosophy of mind, and styled myself an *anti*naturalist, at least in the sense of 'naturalism' that tends to be operative in philosophy of mind.

Recall that, in the first chapter, I distinguished the operative senses of 'naturalism' in philosophy of science and epistemology from those employed in philosophy of mind (and ethics). In epistemology and philosophy of science, 'naturalism' tends to signify an approach that rejects aprioristic armchair methods in favor of approaches that are closely informed by, and perhaps even continuous with, the sciences. In this sense, my project *is* "naturalistic." Cognitive Pluralism, as I have developed it, is a kind of paradigm case of a view of the mind that is driven by evidence taken from a number of the sciences of the mind. And I regard its plausibility as being beholden to further evidence from those fields. As an account of how the mind knows things, it is thus "naturalistic" in much the sense that, say, Goldman's work is "naturalistic" and, say, standard analytic discussions of the Gettier problem are not. Likewise, as a philosophical view about scientific modeling, it *is* an example of cognitivist philosophy of science, and is again driven by, and beholden to, theories of cognition arising from the sciences themselves. It is thus an example of "naturalistic" philosophy of science. I suspect that this is the sort of thing that Millikan had in mind in regarding my project as "naturalistic" as well: that is, that it is pursued in close conjunction with studies of the sciences of the mind.

But 'naturalism' tends to have a very different meaning in philosophy of mind. In chapter 1, I characterized this sort of "naturalism" in terms of a general schema, supplemented by a caveat:

> *Naturalism—a General Schema*: Naturalism about domain D is the view that all features of D are to be accommodated within the framework of nature as it is understood by the natural sciences.
> **Caveat:** A naturalistic theory cannot be one that
> (a) posits the existence of supernatural entities (such as God, angels or immaterial souls), or
> (b) adopts a metaphysical stance in which the ontology of the natural sciences is not fundamental (e.g., transcendental idealism, pragmatism).

Does Cognitive Pluralism hold that mental phenomena can be "accommodated within the framework of nature as it is understood by the natural sciences"? I think the best answer to this is *no*, for both direct and indirect reasons. The direct reason is that it, like other nonreductionist theories, denies that *everything* about the mind can be understood in nonmental terms. The indirect reason is that, as a pluralist theory, it calls into question the very assumption that there is a single, unitary thing that might plausibly be called "*the* framework of nature as it is understood by the natural sciences." If '*the* framework' implies a unitary, all-encompassing model, I deny that there is any such thing, even in such areas as physics and chemistry, much less biology.

Of course, one might decide to take this in the opposite direction, and to hold that what scientific pluralism shows is that we need to reconceive "the natural world" and how "it" is understood by the natural sciences. In the process, the whole distinction between "the mental" and "the natural" might

no longer be tenable. (I think this is akin to the rationale Chalmers employs in describing his dualist account as "naturalistic.") If "the natural" is a domain constituted by things amenable to scientific investigation, then the successes of the sciences of cognition might, by themselves, and without recourse to reduction, bring the mind within the scope of "the natural." This might involve a rejection of the Enlightenment use of the word 'Nature' as a name for a single unified mechanistic system. But perhaps we should merely say "Hurrah!" and so much the worse for the Enlightenment reification of "Nature."

It is far less clear that the Caveat is so easily handled. Cognitive Pluralism does not *require* us to posit nonmaterial entities like Cartesian souls. But it does not debar us from positing them, either. Other "supernatural" entities—God, angels, transcendent moral principles—have not been discussed at all, but I think that these are at least *compatible with* Cognitive Pluralism, though again perhaps not required by it. More problematic still is the fact that the Caveat rules out accounts that do not take the "natural world" as *fundamental*, but cash it out in terms of other things, like the acts of minds or practical and social relations. On my characterization, Cognitive Pluralism, like Idealism and Pragmatism, is a paradigmatically nonnaturalist view.

Some readers will take issue with this last implication. In particular, there is a Pragmatist tradition from William James to Owen Flanagan that styles itself "naturalistic" as well. James and Flanagan seem to mean in particular by this that they avail themselves of the explanatory resources of evolutionary theory in their accounts of the mind. Cognitive Pluralism is free to do so as well, as indeed I did in chapter 8. To my mind, however, this confuses the epistemological and metaphysical notions of "naturalism." My account is "naturalistic" in the epistemologist's sense of making use of scientific explanations (including those from outside the directly cognitive sciences) to explain things about the mind. But these are only partial explanations. What I deny is that these provide *complete* explanations of all mental phenomena, and hence do not underwrite the metaphysical conclusion that the mind is *nothing but* a collection of processes of the sorts studied by the natural sciences. We are natural beings, *among other things*. But some of those other things cannot be explained by appeal to the natural sciences.

Of course, as my account is pluralist rather than dualist, I hold that similar things are true of the relations between things we call "natural sciences." We do not have grounds for saying that biological phenomena are "nothing but" physical phenomena, either. And again, this may force us to revise our usage of "the natural sciences" so that it clearly does not imply that there is a unified *realm* called "Nature" that is exhaustively described by a single theory. More fundamentally, the cognitivist side of Cognitive Pluralism, as developed in chapter 9, entails that there is an important sense in which physical, chemical, and biological processes are *not* "more fundamental" than the mind, as our constitution of these domains depends on the cognitive architecture of the (human) mind.

Of course, there is a level at which the decision about how to use the words 'nature', 'natural', 'naturalism', and 'naturalistic' is a *rhetorical* choice. But that

does not mean that the choice is arbitrary. One *could* choose to reposition these terms in light of a theory like Cognitive Pluralism so that they do more useful work. This, I think, is the strategy that my colleague Joseph Rouse (2003) adopted in *How Scientific Practices Matter*. He started out on this project viewing his own social view of scientific knowledge as antinaturalistic, but in the end decided to co-opt the word 'naturalism' rather than demonize it. In epistemology and philosophy of science, this is probably a good strategy. It is close to the usage already at work in those fields; moreover, linking epistemology and philosophy of science to case studies in the relevant sciences is a beneficial strategy in producing better philosophical accounts.

I think the rhetorical situation is different in philosophy of mind. In the professional literature, and more strikingly in popularized accounts, the words 'naturalism' and 'reductionism' have been strongly linked to a deflationist agenda, which paints a picture of the human person as nothing but a machine or an organism that is exhaustively determined by its physical, biological, and neural properties. At least in the popular imagination, this has been linked with the view that the things we have most deeply assumed to be most fundamental to our nature as human persons—intrinsic worth, moral accountability, free will, and the prospects of survival of death—are precluded by the scientific picture of the world. Of course, even many materialist philosophers have been at pains to deny these conclusions. But their arguments have arguably done less to penetrate the popular consciousness than opposing views expressed by scientists like Skinner, Sagan, Dawkins, Wilson, or Crick.

For these reasons, the ordinary reader, and even the professional philosopher, might be inclined to feel that our core identity as human persons is threatened by reductionism and naturalism. Indeed, I think that they are *right* that *certain* philosophical interpretations of science—particularly reductionism and determinism—*do* threaten our picture of ourselves, particularly as free agents. I think that these threats are *unsuccessful*, in that the philosophical views in questions draw the wrong conclusions from the science, and that the folk are entitled to be reassured. The question, then, is whether this is best accomplished at a rhetorical level by *repositioning* key terms like 'naturalism' and 'reductionism', or by allowing them to stand for legitimately threatening views and arguing forcefully against them. I tend to think that the task of reeducating the public to hear something more subtle and sophisticated when they hear the words 'reduction' and 'naturalization' is very likely impossible to accomplish. And so, instead of co-opting the terms, I seek to bury them. The sciences ought not to be handcuffed so that they say only things we want to hear. But neither ought they to be used to suggest *false* things that are also *harmful*. I consider reductionism, and the kind of ''naturalism'' I have discussed here, to be both false and harmful. I am not sure what I would do if I thought them harmful but *true*, but fortunately that is not the situation I find myself in.

It is possible that this is a point at which the most reasonable strategy to pursue as philosophers of science and the most responsible strategy to pursue as public intellectuals may differ. The attempts by Chalmers and Rouse to

broaden and rework the word 'naturalism', and Bickle's attempt to revivify a more limited use of 'reduction' more in keeping with scientific usage, are well taken. That is, they seem like reasonable moves within a particular academic sphere of discourse. Their intended core readership can be expected to catch on to the subtle rhetorical moves employed in modifying the use of these as technical terms. But academic discussions have a habit of filtering down into the popular press and the public consciousness in ways that are no longer in the control of the authors. The educated nonspecialist who picks up such books at her local bookstore is likely to assume that the author means by these words what *she* has always meant by them, and to draw conclusions very different from those the authors might endorse, especially if the same words are used, much less carefully, by writers like Crick and Wilson in books that command a great deal more shelf space. (Wilson, for example, seems to use stronger and weaker notions of 'reduction' without realizing that they are different. And Crick slips between calling the view that we are nothing but the molecules in our brains *an* "astonishing *hypothesis*" and "*the* scientific *view*.")

My own rhetorical choice has therefore been to go in the opposite direction, and to use words like 'reduction' and 'naturalization' in ways that will seem familiar to both specialists and nonspecialists alike, even though the specialist, at least, may be aware of alternative, and better, usages, and for me to find other ways of talking about ideas that I find both plausible and harmless. The specialist will be able to follow my distinctions, and the nonspecialist may be edified, and perhaps relieved as well.

Of course, both cognitivism and pluralism may strike specialists and nonspecialists alike as unwelcome and threatening in their own ways. They may, for example, suggest to some the view that there is no objective world, or that all interpretations are equally good. I *hope* that my sketchy and initial development of Cognitive Pluralism in this book has provided ample evidence that I do not hold these views. General relativity and quantum mechanics I regard as equally good, even if there are special circumstances in which they generate inconsistent results in combination; Copernican and Ptolemaic cosmology are not equally good. But a more thorough discussion will have to await another, and very different, sort of book devoted to a more thorough exploration of Cognitive Pluralism.

This page intentionally left blank

Notes

CHAPTER 1

1. Other nonphilosophical usages include "An expert in or student of natural history; a person who has a special interest in or makes a special study of plants or animals; (in later use) *esp.* an amateur concerned more with observation than with experiment," or "A creative artist who aims at close representation of nature or reality."

2. There are self-styled naturalists, particularly in the Pragmatist tradition, for whose work the application of this Caveat would be problematic. William James, in particular, styled himself a naturalist in the specifically Darwinian sense of that word, but would have rejected the general project of fundamental ontology, physicalist or otherwise.

3. This is clearly the intent of E. Nagel (1961), for example, and of Descartes's scientific "proofs" in *The World* and *Principles of Philosophy*.

4. In vision science, "psychophysical" laws are generally taken to be those of what Fechner called "outer" psychophysics, dealing with relationships of stimuli and subjective percepts. When Davidson and Chalmers speak of "psycho-physical" laws, they are speaking of relations between brain states and subjective mental states, including percepts.

CHAPTER 2

1. Jackson (1982) proposes the following thought experiment: Imagine that, several decades down the road, when by happy chance and hard work we have discovered all there is to know about the neuroscience of vision, there lives a talented neuroscientist named Mary. Mary, having mastered her chosen field, knows all that there is to know about the neuroscience underlying vision. However, there is a catch to the story. Mary has lived all her life in a room in which she is exposed only to objects that are black or white or some shade of gray. She has never seen red or any other color. Her clothing, her furniture,

her books, her very skin are lacking in chromatic pigment. As a result, she knows all there is to know about the brain processes that go on when one sees, say, red; but she has not seen red herself. Then one day she is shown a red object. She has a new experience: she sees red—or, as philosophers like to say, she experiences red qualia—for the first time. Jackson now bids us ask the question, *Does Mary thereby discover something new that she could not have known before?* The answer that he urges upon us—and a rather intuitive answer it is—is that Mary *does* learn something new: she learns what it is like to see red. (Alternatively, she learns what red looks like, she learns what it is like to experience red qualia, etc.) Jackson then pushes this intuitive answer in two ways: first, because Mary could not have known what it was like to see red without actually experiencing it, it is impossible to *explain* the specific character of red qualia from her previous knowledge base—which was the whole of a completed neuroscience of vision. Second, he claims that she now knows of a new *property*—what it is like to see red—which seems to be a property that slipped the net of neuroscience.

2. Thomas Nagel's provocative "What Is It Like to Be a Bat?" (1974), as the title suggests, involves a thought experiment in which the reader is urged to consider what it would be like to be a bat—and, more specifically, to sense the world through the bat's faculty of echolocation, in which it emits a high-pitched sound and then locates objects by the way they reflect back that sound. In large measure, the point of asking us to imagine this is that we find that it is something we are unable to imagine. We simply do not have the bat's faculties for representing echolocation information, and anything we do using our own faculties (say, imagining high-pitched sounds and picturing outlines of objects at the same time) is no more the same thing than it would be for a blind person to "imagine" vision by reflecting on how things feel to the touch.

3. Searle's (1980) Chinese room thought experiment was posed as a response to Turing's suggestion that a computer that could, by following a program, simulate human linguistic competence to the satisfaction of people "conversing" with it through a teletype, would have demonstrated that it possessed sufficient conditions to be deemed a thinking thing. Searle suggests that we replace the computer and program in this scenario with a monolingual English-speaking human following an algorithmic program in the form of a rule book that tells him what to do when various Chinese symbols are passed to him inside of a room closed off from the outside. Anything that can be rendered as a computer program can be put into this rule book, and so the situation is precisely the same as a computer given the Turing Test in Chinese. Searle's point is this: even if we spot Turing the assumption that linguistic competence can be simulated through this sort of algorithmic rule-following, it is plain that the man in the room does not *understand* Chinese, despite all appearances to the people outside the room. And because understanding is a necessary condition for thought, program-following cannot be a sufficient condition for thought or intentionality.

4. Jackson's views on this have changed in recent years. He is now a proponent of nonreductive materialism.

5. Carnap's axiomatic vision of scientific unification is not the only way the phrase "unity of science" has been taken. Indeed, Otto Neurath, the editor of the *Encyclopaedia of Unified Science*, seemed to view the "unity" of the sciences more on the model of the unity of an encyclopaedia, with separate articles and liberal cross-referencing (Cat et al. 1996). Ian Hacking (1996) has also usefully distinguished "unity as singleness" and "unity as harmonious integration."

6. For Platonists, this was largely a consequence of the view, derived from the *Timaeus*, that matter is a chaotic and incomprehensible substratum to which the Forms are applied by the World-Soul. Physical objects result from the application of

the Forms, which are fully intelligible and completely real, to matter, which is unintelligible; they stand between Forms and matter in both degree of being and intelligibility. Aristotle's more empiricist epistemology led to the conclusion that we cannot perfectly discover the underlying entelechies of material things, and thus are incapable of fully understanding them.

7. I am wary of the expression 'conceptual necessity' because it seems to mix together issues that are primarily metaphysical with those that are primarily epistemological. This creates an ambiguity: is 'conceptual necessity' (a) a type of *necessity* (a metaphysical feature) or (b) a matter of what can be understood or inferred on conceptual grounds (an epistemic feature)? To my mind, what is really aimed at is simply that subset of *metaphysical* necessities that can be known through conceptually adequate explanation. My way of putting the issue separates the metaphysical and the epistemological problems.

8. New Semantics can also be given nonessentialist and nonrealist interpretations.

9. I have never been completely satisfied that 'water' *is* really a filler term, nor indeed (a) that it need function in the same way for all English speakers, or even (b) that the question of whether it is a role term has a fully determinate answer for every individual English speaker. When one presents the Kripke or Putnam analysis to a roomful of people for the first time, there are almost inevitably those whose response is to say that there is water on Twin-Earth too. It is not obvious that this represents any confusion on their part. They may be reporting, honestly and accurately, the semantics of their own use of the word 'water' as a role term. If they respond differently after a course in philosophy of language or metaphysics, this may represent a *change* in how they use the term rather than a better self-understanding. This case seems even more plausible for prescientific people. I find it unlikely that there was something about Shakespeare's use of the word 'water' that determined that it would track specifically a molecular kind. Likewise, there are people who respond to the Twin-Earth cases with a kind of bafflement or indecision about what should be said about them. This could quite plausibly be a symptom of a real indeterminacy in their previous usage. For them, adopting the orthodox philosophical analysis may serve more as a *regimentation* of previous usage than an analysis thereof.

Chapter 3

1. A possible second problem with the derivation of the Boyle-Charles laws is independent of Garfinkel's criticism. This second problem consists in the fact that, even if we allow that the derivation goes through, it does not account for all of the features of the thermodynamic system *without remainder*. In particular, there is a crucial temporal asymmetry in thermodynamics: entropy is a one-directional feature and is not temporally reversible. But the properties of particle collisions in Newtonian mechanics are all temporally symmetrical and cannot be used to derive a temporal asymmetry. The asymmetry is standardly understood to be explained by assumptions about the initial conditions of the system. This, however, is an additional assumption, though arguably not as problematic as the one Garfinkel points out.

2. Van Fraassen (1988) has positioned himself as an empiricist in the debate over scientific realism but gives a pragmatic/erotetic account of the nature of explanation. On this particular issue of scientific realism, empiricism and pragmatism make for easy allies. Kitcher (1993) assesses the rationality of science in terms of whether it satisfies criteria of good design relative to collective goals—a social Pragmatist view. But his unificationist account of relations between theories is also compatible with an

empiricist reading, in which progress consists in a succession of ever broader and more powerful empirical generalizations.

3. Cartwright's more recent work has such causal notions as the "capacities" of objects. However, one can arguably see Pragmatist themes in Cartwright's restriction of the truth of scientific claims to particular contexts and experimental setups.

4. In both the eighteenth and the nineteenth centuries, there were also influential alternatives on the scene. Self-styled "Newtonian" Empiricists like Locke and Hume saw little hope for extending the model of mathematical deduction to the natural sciences, nor did nineteenth-century Positivists like Mach.

Chapter 4

1. In the seventeenth century, and even into the eighteenth, '*scientia*' was restricted to knowledge that was known either directly and indubitably on first principles, or things validly deduced from such knowledge. Hence the oddity of Locke's and even Hume's qualms about applying this honorific (or its English translation as 'knowledge') to Newtonian science.

2. Newton started his career as a Cartesian, and hence assumed that all of physics was confined to mechanical contact interactions. His discovery of an independent gravitational force was thus a kind of blow to a received understanding of scientific unity, and the more controversial as it required action at a distance (for which Leibniz and others accused him of reverting to the "occult forces" for which Mechanists had excoriated Scholastic science). Newton suggests in the Scholia that there might be any number of additional forces yet undiscovered. He was much more careful than Descartes in his standards for what counted as a successful explanation, and hence more realistic than Descartes in recognizing things that were still unexplained by his theory. He was also very cautious about "hypotheses" going beyond the evidence: for example, in the famous passage from the Scholia to the second edition of the *Principia* quoted earlier, he goes so far as to acknowledge that his mechanics requires the postulation of a gravitational *force* but refuses to postulate a *mechanism* through which it might operate. He also seems to have been quite the pluralist with respect to possible sources of knowledge, as he spent a great portion of his life with alchemical experiments, studies of biblical prophecy, and studies of ancient mystical traditions such as Orphism. His philosophical followers, such as Locke and Hume, however, were not immersed in the science and tended (as philosophers are wont to do) to see only the successes of the *Principia* and *Optics* and to ignore the empirical caution and methodological pluralism characteristic of Newton himself. (Some of this, however, may be due to Newton's own secrecy about his alchemical and theological work. The former was a secretive tradition, and the latter involved heretical Arian tendencies that would have landed even a man of Newton's stature in deep trouble in 1700.)

3. For a broader account of the motivations, see M. Wilson (2006).

4. William Bechtel, in commenting on the manuscript of this book, pointed out to me that identity claims have implications that causal claims and correlations do not. For example, identity claims have two-way implications, whereas causation is a one-way relation, and correlation is a statistical notion and is only compatible with identity when the correlation is perfect. Thus, for example, the claim that Mendel's factors are parts of chromosomes committed the Morgan school to a research programme that weaker claims could not have. Bechtel is certainly correct about the *commitments* involved in different metaphysical interpretations. My point, however, is about the *evidence* for these claims. In the initial discovery of correlations between variables, we do not have

evidence to adjudicate between competing metaphysical interpretations. Adopting a strong interpretation (e.g., identity) may later yield results that are incompatible with weaker interpretations, but assessing whether it does so in a given case requires a great deal of attention to detail (which is to say, exactly the sort of philosophical treatment that Bechtel excels at).

5. As I've argued elsewhere in the book, it will not do to interpret the laws as universally quantified claims hedged by *ceteris paribus* clauses. If so interpreted, most laws would turn out to be false, and indeed to have no true substitution instances. Nor is it faithful to scientific practice to view them as laws that are true, but true of idealized worlds. Scientists view laws as making true claims about the real world.

Chapter 6

1. These formulations utilize a notion of "time *t*" that may be strictly incompatible with relativistic physics' denial of absolute simultaneity. However, in the case of local supervenience and causal closure, this is presumably innocent, so long as the relevant basal and causal facts lie within the backward light cone of the phenomenon in question. In such cases, we may either assume that frame of reference is unimportant to the characterization of the phenomena to be related, or else assume the frame of reference of the supervening phenomena or effects. Notions of "global supervenience at a time *t*" are arguably more problematic in this regard, though versions of global supervenience that include the entire history of the universe might escape this problem.

2. This way of construing events is controversial. If events are individuated by their mode of presentation, we are debarred from assuming token identity of events with different descriptions.

3. The individuation of events is a potentially thorny issue here. One might embrace a strict identity logic for objects while insisting on a relative identity logic for events. The case for doing so is less compelling, in my view, than the case for a relative identity logic for actions, but it is a sort of metaphysical issue that is sometimes unjustly ignored in philosophy of mind.

4. Of course, there are a number of deep issues under the surface here—issues that have to do with how we construe events, objects, and identity. For example, if we assume a relative identity logic, there is no such thing as simple numerical identity: for the relative-identity theorist, saying that X and Y are identical is always (implicitly) saying that they are "the same P," where 'P' stands in for some kind-term. (And generic sortals like 'object' and 'event' are not legitimate kind-terms in relative identity logic.) Likewise, if mode of presentation is essential to the individuation of events, then "this pain" and "this C-fiber firing" cannot be the same event, because they are individuated through different modes of presentation. Token physicalism assumes that events, like individual objects, are individually of mode of presentation. I think there is something right about this assumption, but it is not uncontroversial.

Chapter 7

1. Rosenberg (1994), for example, admits a pluralism in biological categories, but takes the view that this makes them only instrumentally useful, rather than *true* in the full sense applicable to claims in chemistry and physics.

2. The relation between cognitivist and Pragmatist themes ultimately raises some important questions about the relations of minds and practices, and indeed where we draw the boundaries of "the mind." For example, while the design of the human

brain no doubt places constraints upon the types of thinking available to human minds, it also seems clear that social practices (e.g., language games) and material practices (e.g., storing information in writing or in computer media, or calculation techniques, of both the pen and paper type and those performed by machines) can expand both the kinds and the power of thinking available to us. When one is "thinking" with the help of pen and paper, or a computer, or a team of other researchers with different specialized skills, we have situations that invite further questions about whether our notion of "mind" should include such extended and distributed practices *within* it. (For example, some "mental" abilities may be best understood as part of an extended phenotype of groups of humans.) These issues, however, go beyond the scope of this book.

3. Eric Watkins has told me in conversation that Kant's Third Critique treats biology as being in important ways discontinuous with physics. Kant was not, of course, a party to twentieth-century conversations about the "unity of science," but it is plausible to suppose he would have thought that at least biology cannot be unified with physics. I also tend to think that the First Critique's discussion of "the world" as one of the dialectical illusions of Reason suggests that Kant would have been suspicious of any dreams of uniting all our knowledge of the world in the form of a single system, even if Reason has an innate drive to work toward such an end.

4. While such abstraction is often the result of starting with the complexity of the real world and abstract*ing* away from some features of it, abstract modeling can also take on a fruitful life of its own, in which attention to models that are abstract—but not a result of abstract*ing*—are used to generate hypotheses. See the essays in Morgan and Morrison (1999).

5. There are, of course, dangers in using any such "representationalist" metaphors, such as "cognitive lenses" or "representational systems." Taking them as unidealized bedrock truths, rather than as metaphors that are useful in one context but potentially harmful in another, leaves us at risk of the skepticism that has often resulted from cognition as accomplished through some intermediary. I explored some such dangers in Horst (1996), and my colleague Joe Rouse has pointed out to me that Hegel argued long ago that such metaphors lead to skepticism. However, as a *pluralist*, I contend that such metaphors are useful in some contexts, such as disabusing us of the implications of other metaphors, such as the "mirror of nature." They are also arguably necessary whenever we turn to certain types of questions in naturalistic epistemology, such as how a way of thinking about the world captures, or screens out, features of the world.

6. I would take Giere's (1988, 1999, 2004) view that models *resemble* the real world as falling within this category. While I find Giere's views on the whole sympathetic to my own, on this point we disagree. I think we can make sense of the view that models represent aptly, and hence that statements made within the vocabulary of a model can say things that are true, if we view models as idealized.

7. A paper by Eric Winsberg (forthcoming) explores examples of this in nanotechnology, where he claims one not only needs to employ three inconsistent models, but to appeal to clearly fictitious "molecules" in the process.

8. The type of "independence" assumed between relativistic gravitation and quantum mechanics may be importantly different from that assumed between forces in classical mechanics, as the latter assumes a common domain of classical bodies to which various classical forces can apply.

9. This, of course, is one of the concerns that leads some physicists to seek a theory of quantum gravity. If such a theory is to be had, this problem may be tractable after all. However, for the moment, I regard "quantum gravity" as an item on the wish list of

some physicists, rather than as a fixed point in contemporary physics from which we are entitled to make inferences.

10. This possibility is in part dependent on what is the proper interpretation of laws. For example, advocates of antirealist accounts of laws would likely assert that this scenario rests on a philosophical misunderstanding.

11. The notion of "context" here is one that builds in the pragmatic use and goals.

12. I had initially assumed that this was a straightforward consequence of General Relativity. In e-mails with mathematical physicist Tom Ilmanen, however, I learned that, though this was Einstein's own view, it is possible to understand General Relativity as permitting a flat, massless space.

13. Rouse (1987, chapter 5) makes a related suggestion, to the effect that 'truth' applies to statement tokens rather than statement types, and that the *use* of models includes an understanding of how to handle the idealizations in different cases. Whether one construes the notion of 'model' so as to include such understanding of how and when to apply it, or supplements a more limited notion of 'model' with such understanding, one must somehow build this pragmatic know-how into one's conception of the use of models in science.

14. Paralogism is a form of logical fallacy in which a term shared between two premises is ambiguous, and different legs of the ambiguity are operative in the different premises. For example:

P1. Aunt Polly deposited $100 at a bank.
P2. A bank is the side of a river.
P3. Aunt Polly deposited $100 in the side of a river.

15. Cartwright sometimes goes too far, in my view, in claiming that experiments performed in such regimented situations give us literally no reason to believe that, say, gravitation behaves the same way outside of those situations (though she gives a more balanced view in chapter 1 of *The Dappled World* [1999]). She is right that it gives no reason to think that the real-world *motions* of objects will be the same, since other forces are at work. But I contend that we have good reason to think that gravitational *force* is operative in the same ways. However, Cartwright does not believe in component forces.

Chapter 8

1. Of course, anything that can be carried out by program-driven central processing can also be done by a dedicated circuit board. There was a time, for example, when there were production-model machines dedicated exclusively to tasks such as word processing. A computational task can generally be done more quickly (often by several orders of magnitude) by a dedicated piece of hardware, but a program can be edited and updated much more easily, without replacing hardware, and the same program can often be run on newer and faster hardware. As processing speed of computers increased exponentially, it became an increasingly attractive design choice to minimize the number of special-purpose modules in production-model computers and to implement more and more functionality through programming.

2. One might reasonably credit Gall (the founder of phrenology) with anticipations of the general *thesis* that particular cognitive abilities and traits are tied to particular regions of the brain, even though the phrenological method was flawed, and Gall's own conjectures as to how cognitive traits map onto brain areas was deeply mistaken.

3. This is a slight simplification. The color, motion, and form pathways are all subsystems within one of three visual "systems" that seem to be of different phylogenetic ages: the Thalamofugal, the Tectofugal, and the Accessory Optic System. The Accessory Optic System, which plays a role in self-motion and gaze stabilization, does not pass through the visual cortex. The Tectofugal System plays roles in visual orientation and spatial attention. It contains circuits in the Superior Colliculus and Pulvinar Nucleus, which feed into V2. It is the Thalamofugal pathway that contains subpaths for form, color, and motion.

4. Best documented are differences in lateralization of linguistic areas corresponding to handedness. Right-hand-dominant individuals generally have activity during language-processing production in the left hemisphere of their brain. This is reversed in about 50 percent of left-handed individuals.

5. Related studies showing that the visual cortex of blind subjects is active during the reading of Braille strikes me as unsurprising and unproblematic. Given that there are feed-forward connections from tactile inputs to the visual cortex even in the brains of sighted individuals, it is unsurprising that this area would be available for processing of tactile information. Moreover, there could well be a particular kind of fit between the structure of areas of the visual cortex and problems dealing with spatial patterns or even symbols, making it less surprising that this *particular* sort of activity might involve visual areas. That is, those areas may be utilized, not because they are now "tactile areas," but because they are areas applied in sighted and unsighted subjects to particular *abstract spatial* problems. This of course still serves as evidence for a more complicated relationship between functional and anatomical typing of cortical areas, as it would imply that such areas should be viewed as "amodal" rather than "visual." However, what these experiments get at strikes me as being altogether different from the blindfold experiments.

6. It is possible that we should also wish to speak of subconceptual systems in simple organisms (and ones that are still found in human beings) as "models" as well. The same considerations apply, *mutatis mutandis*, to these.

7. In saying they are "optimized" for pragmatic purposes, I do not mean that they perform *optimally*. I do not take the view that evolution will hit on optimal solutions, only satisficing ones. They involve attraction to local maxima, but not necessarily to the global maximum, as initial trajectories may make this impossible, as may constraints of the biological nature of the organism, or indeed of the laws of physics. A truly *optimal* animal, for example, would be a perpetual-motion machine, but physics does not cooperate with this outcome.

8. I should note, however, that theological tradition tends to view direct apprehension of the real natures of created things as something lost in the Fall, leaving us with a combination of animal mechanisms and computational reasoning—a kind of amalgam of monkey and Macintosh. It is only in spiritual enlightenment, not in science, that the saints are supposed to regain something more, and then only by the infusion of the Holy Spirit.

9. In terms of what Kant did or did not get right, it is important to note that his claim is restricted to a claim about how we can represent things *in Sensibility*. This is compatible with our being able to represent non-Euclidean geometries mathematically, using Understanding, though this would present a problem for Kant if the concepts thus employed could not be fulfilled in a sensuous intuition.

Chapter 9

1. It is not clear that Descartes was perfectly consistent on this point. In his early works on mechanical psychology, particularly *Treatise on Man*, but also cognate sections

of the *Discourse on Method*, Descartes attributes a surprising array of psychological capacities to "that machine," the body, including such things as memory and sensation. Only reasoning and language are argued to be such that they cannot be mechanical, though one might presume that Descartes would add free will to the list if pressed. It is hard to render this consistent with the position of his midcareer works, particularly the *Meditations*, that mind and body are completely distinct, unless perhaps by assuming that Descartes really meant that *only* reason, language, and the will were to be attributed to the soul, and that sensations were *not* modifications of the soul, but of the body. This, however, is in tension with the very natural assumption that the modifications of the soul are whatever can be known indubitably under radical doubt. In his later years, beginning in correspondence with Princess Elizabeth and culminating in his final work, *The Passions of the Soul*, Descartes took what appears to be a new view: namely, that there are some modifications (the passions) that must be attributed *jointly* to body and soul. This strikes me as a significant modification of the metaphysics of the *Meditations* and even the *Principles*.

2. This example is based on one taken from a discussion paper by Aaron Edidin presented to a reading group at Notre Dame in or around 1984. I am not sure if a descendent of this paper was published.

3. As argued elsewhere in the text, I do not believe that a proper interpretation of laws implies determinism. The illusion that they do so is partially a consequence of misinterpreting laws as making universal claims (in the most plausible versions, modally strengthened) about the behavior of real-world objects. This is a mistake Kant makes in linking the category of cause and effect to a form of statement in first-order logic.

4. I am thinking here of cases in which a subject is psychologically unable to do as duty demands, whether because of weakness of the will, mental illness, or lack of the psychological skills necessary to recognize or act on duty. The question is how to interpret such cases from a deontological perspective. One possibility is that such persons still have the duty, but cannot act on it (or in some cases even comprehend it), violating Kant's principle of "ought implies can." A second possibility is that such persons do not have the duty for reasons fully understandable *within* the deontological framework. If "ought implies can" is part of the deontological model, than we can apply modus tollens and derive the result that if a person literally cannot do A, then she cannot be obliged to do A. This would be problematic for some deontologists (including, I think, for Kant) as it would make it too easy to escape the demands of duty. If one becomes sufficiently a wastrel, one is no longer obliged by the duties one has rendered oneself incapable of fulfilling. (The position is more attractive with respect to those whose development has prevented them from being capable of grasping duty in general, or particular duties, though at the expense of not counting them as Kantian moral agents.) A third possibility, which I favor, is to view claims of duty as self-contained in a deontological model that does not contain within it principles for when it is and is not applicable. The fact that a given person suffers compulsions or cognitive limitations that prevent her from acting in accordance with duty, or recognizing duty, stands outside of the sphere of the deontological model, and what we should say about such cases must be addressed in some other fashion. In some cases, such as weakness of will, I would be inclined to say that "ought implies can" does not hold good: one can both be obliged to do something and be psychologically incapable of doing it. This certainly seems to be an experience that has some phenomenological grounding in that we sometimes experience ourselves as both acknowledging a duty and finding ourselves unable to perform it. (St. Paul vividly describes such situations in chapter 7 of his letter

to the Romans, and it is a central tenet of Christian theology. It is also a central principle of twelve-step programs like Alcoholics Anonymous.) On the other hand, cases of cognitive defect may place a person in a category to which deontological principles, or at least some particular deontological principles, do not properly apply. It may make no more sense to ascribe duties to persons so profoundly retarded that their mental age is that of an infant than it does to ascribe duties to actual infants or to nonhuman animals.

Bibliography

Abelson, R. P. 1973. The Structure of Belief Systems. In *Computer Models of Thought*, edited by R. C. Schank and K. M. Colby. San Francisco: Freeman.

Abelson, R. P., and C. M. Reich. 1969. Implication Molecules: A Method for Extracting Meaning from Input Sentences. Paper presented at International Joint Conferences on AI, Washington, DC.

Anderson, John R. 1980. On the Merits of ACT and Information-Processing Psychology: A Response to Wexler's Review. *Cognition* 8: 73–88.

Anderson, John R., and Gordon H. Bower. 1973. *Human Semantic Memory*. Washington, DC: Winston.

Anderson, Michael. Forthcoming. The Massive Redeployment Hypothesis and the Functional Topography of the Brain. *Philosophical Psychology*.

Austin, J. L. 1962. *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Baker, Lynne Rudder. 1995. *Explaining Attitudes*. New York: Cambridge University Press.

Bechtel, William. 1983. Forms of Organization and the Incompletability of Science. In *The Limits of Lawfulness*, edited by N. Rescher. Lanham, MD: University Press of America.

——— . 1984. Reconceptualizations and Interfield Connections: The Discovery of the Link Between Vitamins and Coenzymes. *Philosophy of Science* 51: 265–92.

——— . 2006. *Discovering Cell Mechanisms: The Creation of the Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Belot, G., and J. Earman. 1997. Chaos out of Order: Quantum Mechanics, the Correspondence Principle and Chaos. *Studies in the History and Philosophy of Modern Physics* 2: 147–82.

Bermudez, José. 2003. *Thinking Without Words*. Oxford: Oxford University Press.
Bickle, John. 1998. *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.
—— . 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.
Block, Ned. 1978/1980. Troubles with Functionalism. In *Readings in the Philosophy of Psychology*, edited by N. Block. Cambridge, MA: MIT Press. (Originally published in *Minnesota Studies in the Philosophy of Science* 9: 261–325.)
Block, Ned, and Robert Stalnaker. 1999. Conceptual Analysis, Dualism, and the Explanatory Gap. *The Philosophical Review* 108:1–46.
Boveri, Theodor. 1903. Uber die Konstitution der Chromatischen Kernsubstanz. *Verh Zool Ges Wurtzburg* 35.
Brandom, Robert. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.
Brandon, Robert N. 1985. Grene on Mechanism and Reductionism: More Than Just a Side Issue. *PSA* 1984 (2): 345–53.
Brodmann, Korbinian. 1909. *Vergleichende Lokalisationslehre der Grosshirnrinde*. Leipzig: Barth.
Burge, Tyler. 1979. Individualism and the Mental. *Midwest Studies in Philosophy* 4: 73–122.
Callebaut, Werner. 1993. *Taking the Naturalistic Turn, or, How Real Philosophy of Science Is Done*. Chicago: University of Chicago Press.
Campbell, D. T. 1974. "Downward Causation" in Hierarchically Organized Biological Systems. In *Studies in the Philosophy of Biology*, edited by F. J. Ayala and T. Dobzhansky. London: Macmillan.
Carey, Susan. 1998. Knowledge of Number: Its Evolution and Ontogenesis. *Science* 242: 641–42.
Carruthers, Peter. 2002. The Cognitive Functions of Language. *Behavioral and Brain Sciences* 25 (6): 657–74.
Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. New York: Oxford University Press.
—— . 1989. *Nature's Capacities and Their Measurement*. Oxford: Clarendon.
—— . 1994. Fundamentalism vs. the Patchwork of Laws. *Proceedings of the Aristotelian Society* 94: 279–82.
—— . 1997. Why Physics? In *The Large, the Small and the Human Mind*, edited by R. Penrose, A. Shimony, N. Cartwright, and S. Hawking. Cambridge: Cambridge University Press.
—— . 1999. *The Dappled World: A Study of the Boundaries of Science*. New York: Cambridge University Press.
Cat, Jordi, Nancy Cartwright, and Hasok Chang. 1996. Otto Neurath: Politics and the Unity of Science. In *The Disunity of Science: Boundaries, Contexts, and Power*, edited by P. Galison and D. Stump. Stanford: Stanford University Press.
Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
Chalmers, David J., and Frank Jackson. 2001. Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110 (3): 315–60.
Changeux, J.-P. 1985. *Neuronal Man: The Biology of Mind*. Oxford: Oxford University Press.
Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
—— . 1966. *Cartesian Linguistics*. New York: Harper and Row.
—— . 1980. *Rules and Representations*. Oxford: Basil Blackwell.
Churchland, Patricia S. 1986. *Neurophilosophy*. Cambridge, MA: MIT Press.
Churchland, Paul M. 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78: 67–90.

Colby, Kenneth M. 1975. *Artificial Paranoia*. New York: Pergamon.
Cosimedes, L., and J. Tooby. 1987. From Evolution to Behavior: Evolutionary Psychology as the Missing Link. In *The Latest on the Best: Essays on Evolution and Optimality*, edited by J. Dupré. Cambridge, MA: MIT Press/Bradford Books.
Cowie, Fiona. 1999. *What's Within: Nativism Reconsidered*. New York: Oxford University Press.
Craver, Carl. 2002. Structures of Scientific Theories. In *The Blackwell Guide to the Philosophy of Science*, edited by P. Machamer and M. Silberstein. Malden, MA: Blackwell.
Crick, Francis. 1993. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Scribner.
Crick, Francis, and Cristoph Koch. 1990. Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences*, no. 2: 263–75.
Cummins, Robert. 1996. *Representations, Targets and Attitudes*. Cambridge, MA: MIT Press.
Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Putnam.
Darden, Lindley, and Nancy Maull. 1977. Interfield Theories. *Philosophy of Science* 44: 43–64.
Davidson, Donald. 1970. Mental Events. In *Experience and Theory*, edited by L. Foster and J. Swanson. Amherst: University of Massachussetts Press.
——— . 1984. Thought and Talk. In *Truth and Interpretation*. Oxford: Clarendon Press.
Davies, Paul C. W. 2004. Emergent Biological Properties and the Computational Properties of the Universe. *Complexity* 10 (2): 11–15.
Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
Dennett, Daniel. 1991. *Consciousness Explained*. Boston: Little, Brown.
Descartes, René. 1985. *The Philosophical Writings of Descartes*. Translated by J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny. 3 vols. Cambridge: Cambridge University Press.
Dretske, Fred. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.
Dreyfus, Hubert. 1979. *What Computers Can't Do*. New York: Harper and Row.
Dupré, John. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.
——— . 2001. *Human Nature and the Limits of Science*. Oxford: Oxford University Press.
Edelman, Gerald M. 1988a. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.
——— . 1988b. *Topobiology: An Introduction to Molecular Embryology*. New York: Basic Books.
Fechner, Gustav T. 1877. *In Sachen der Psychophysik*. Leipzig: Breitkopf & Härtel.
Felleman, Daniel J., and David C. van Essen. 1991. Distributed Hierarchical Processing in the Primate Visual Cortex. *Cerebral Cortex* 1: 1–47.
Fodor, Jerry. 1974. Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese* 28: 97–115.
——— . 1975. *The Language of Thought*. New York: Thomas Crowell.
——— . 1981. *RePresentations*. Cambridge, MA: Bradford Books/MIT Press.
——— . 1983. *Modularity of Mind*. Cambridge, MA: MIT Press.
——— . 1986. Précis of *The Modularity of Mind*. In *Minds, Brains and Computers*, edited by R. Cummins and D. Cummins. London: Blackwell.
——— . 1987. *Psychosemantics*. Cambridge, MA: Bradford Books.
——— . 2000. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.
Freeman, Walter J. 1991. The Physiology of Perception. *Scientific American* 264 (2): 78–85.

Freeman, Walter J. 1995. *Societies of Brains: A Study in the Neuroscience of Love and Hate*. Freeman, NJ: Erlbaum.

Gardner, Howard E. 1999. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

Garfinkel, Alan. 1981/1999. Reductionism. In *The Philosophy of Science*, edited by R. Boyd, P. Gaspar, and J. D. Trout. Cambridge, MA: MIT Press. (Originally published as chapter 2 in *Forms of Explanation*. New Haven: Yale University Press.)

Georgalis, Nicholas. 2006. *The Primacy of the Subjective: Foundations for a Unified Theory of Mind and Language*. Cambridge, MA: MIT Press.

Giere, Ronald N. 1988. *Explaining Science: A Cognitivist Approach*. Chicago: University of Chicago Press.

——— . 1999. *Science Without Laws*. Chicago: University of Chicago Press.

——— . 2004. How Models Are Used to Represent Reality. *Philosophy of Science* 71 (Supplement): S742–52.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

——— . 1992. *Liaisons*. Cambridge, MA: MIT Press.

Gorman, R. Paul, and Terrence J. Sejnowski. 1988. Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets. *Neural Networks* 1: 75–89.

Gould, Stephen J., and Richard C. Lewontin. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptational Programme. *Proceedings of the Royal Society of London B* 205: 581–98.

Grice, Paul. 1957. Meaning. *Philosophical Review* 66: 377–88.

Grossberg, Stephen. 1987. *The Adaptive Brain*. Amsterdam: North Holland.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

——— . 1996. The Disunities of the Sciences. In *The Disunity of Science: Boundaries, Contexts, and Power*, edited by P. Gallison and D. Stump. Stanford: Stanford University Press.

Hayes-Roth, Barbara. 1980. *Human Planning Processes*. Santa Monica, CA: Rand.

Hellman, G. 1999. Reduction (?) of What (?): Comments on L. Sklar's "The Reduction (?) of Thermodynamics to Statistical Mechanics." *Philosophical Studies* 95 (1–2): 200–13.

Hobbs, Jesse. 1993. Naturalism: A Contemporary Shibboleth? Paper presented at NEH Summer Institute on Naturalism, University of Nebraska at Lincoln, June.

Horgan, Terry, and John Tienson. 2002. The Intentionality of Phenomenology and the Phenomenology of Intentionality. In *Philosophy of Mind: Classical and Contemporary Readings*, edited by D. J. Chalmers. New York: Oxford University Press.

Horst, Steven. 1996. *Symbols, Computation and Intentionality: A Critique of the Computational Theory of Mind*. Berkeley: University of California Press.

——— . 1999. Evolutionary Explanation and the Hard Problem of Consciousness. *Journal of Consciousness Studies* 6 (1): 39–48.

——— . 2004. Laws, Mind and Freedom. Manuscript available at http://shorst.Web.Wesleyan.edu.

——— . 2005. Modeling, Localization and the Explanation of Phenomenal Properties. *Synthese* 104 (1): 123–45.

——— . *Laws, Mind and Freedom*. Manuscript available at: //http:shorst.web.wesleyan.edu.

Husserl, Edmund. 1913/1931. *Ideas: General Introduction to Pure Phenomenology*. Translated by W. R. B. Gibson. London: Collier Books. (Originally published as *Ideen au eine reinen Phänomenologie und phänomenologischen Philosophie*. The Hague: Nijhoff.)

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–66.

Jacobson, John. 1997. The Mysteries of Consciousness and the Foundations of the Broad Approach. Undergraduate thesis, Wesleyan University.

Kandel, Eric R., James H. Schwartz, and Eric R. Jessell. 2000. *Principles of Neural Sciences*. New York: McGraw-Hill.

Kant, Immanuel. 1781/1968. *Critique of Pure Reason*. Translated by N. K. Smith. London: Macmillan.

Kim, Jaegwon. 1993. *Supervenience and Mind*. New York: Cambridge University Press.

Kitcher, Philip. 1981. Explanatory Unification. *Philosophy of Science* 48: 507–31.

——— . 1984. 1953 and All That: A Tale of Two Sciences. *Philosophical Review* 93: 335–73.

——— . 1993. *The Advancement of Science*. Oxford: Oxford University Press.

——— . 2003. *In Mendel's Mirror: Philosophical Reflections on Biology*. Oxford: Oxford University Press.

Kramer, Peter. 1993. *Listening to Prozac*. New York: Penguin.

Kripke, Saul. 1972. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lashley, K. S. 1929. *Brain Mechanisms and Intelligence*. New York: Dover.

Laudan, Larry. 1977. *Progress and Its Problems*. Berkeley: University of California Press.

Levine, Joseph. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354–61.

Levins, R. 1968. *Evolution in Changing Environments*. Princeton, NJ: Princeton University Press.

Lewis, David. 1978. Mad Pain and Martian Pain. In *Readings in the Philosophy of Psychology*, edited by N. Block. Cambridge, MA: MIT Press.

——— . 1986. Causation. In *Philosophical Papers*, edited by D. Lewis. Oxford: Oxford University Press.

Lewontin, Richard C. 1983. Biological Determinism. In *Tanner Lectures on Human Values*. Salt Lake City: University of Utah Press.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman.

Mayden, R. L. 1997. A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problem. In *Species: The Units of Biodiversity*, edited by M. A. Claridge, H. A. Dawah, and M. R. Wilson. London: Chapman and Hall.

McGinn, Colin. 1983. *The Subjective View*. Oxford: Clarendon.

——— . 1990. *The Problem of Consciousness*. New York: Blackwell.

——— . 1991. *The Problem of Consciousness: Essays towards a Resolution*. New York: Blackwell.

Meltzoff, A. N., and M. K. Moore. 1977. Imitation of Facial and Manual Gestures by Human Neonates. *Science* 198: 75–78.

Millikan, Ruth. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Minsky, Marvin. 1974. A Framework for Representing Information. MIT-AI Laboratory memo 306. Reprinted 1981 in *Mind Design*, edited by J. Haugeland. Cambridge, MA: MIT Press.

——— . 1985. *The Society of Mind*. New York: Simon & Schuster.

Minsky, Marvin, Push Singh, and Aaron Sloman. 2004. Abstract to "The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence." *AI Magazine* (summer): 113–24.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Morgan, Mary S., and Margaret Morrison, eds. 1999. *Models as Mediators*. Cambridge: Cambridge University Press.

Nagel, Ernest. 1955. Naturalism Reconsidered. *Proceedings and Addresses of the American Philosophical Association* 28: 5–17.

Nagel, Ernest. 1961. *The Structure of Science*. New York: Harcourt, Brace and World.

Nagel, Thomas. 1974. What Is It Like to Be a Bat? *Philosophical Review* 4: 435–50.

—— . 1986. *The View from Nowhere*. New York: Oxford University Press.

Nersessian, Nancy J. 1992. How Do Scientists Think? Capturing the Dynamics of Conceptual Change. In *Cognitive Models of Science*, edited by R. N. Giere. Minneapolis: University of Minnesota Press.

Newell, Alan, and Herbert Simon. 1956. The Logic Theory Machine. *IRE Transactions on Information Theory* 2: 61–79.

—— . 1963. Computers in Psychology. In *Handbook of Mathematical Psychology*, edited by R. D. Luce, R. R. Bush, and E. Galanter. New York: Wiley.

Newton, Isaac. 1713/1962. *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and System of the World*. Translated by A. Motte and F. Cajori. Los Angeles: University of California Press.

Newton-Smith, W. H. 1981. *The Rationality of Science*. Boston: Routledge & Kegan Paul.

Norman, Donald A., and David E. Rumelhart. 1975. *Explorations in Cognition*. San Francisco: Freeman.

Oppenheim, Paul, and Hilary Putnam. 1958. Unity of Science as a Working Hypothesis. In *Concepts, Theories, and the Mind-Body Problem*, edited by H. Feigl, M. Scriven, and G. Maxwell. Minneapolis: University of Minnesota Press.

Papineau, David. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

—— . 2001. The Rise of Physicalism. In *Physicalism and its Discontents*, edited by Carl Gillett and Barry Loewer. Cambridge: Cambridge University Press, pp. 3–36.

—— . 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.

Pascual-Leone, Alvaro, Amir Amedi, Felipe Fregni, and Lofti B. Merabet. 2005. The Plastic Human Brain Cortex. *Annual Review of Neuroscience* 28: 377–401.

Pattee, Howard H., ed. 1973. *Hierarchy Theory: The Challenge of Complexity*. New York: Braziller.

Paulk, Angelique C., and Wulfia Gronenberg. 2005. Color and Motion-Sensitive Cells in the Bee Brain. Paper presented at Program No. 295.17, Society for Neuroscience, Washington, DC.

Place, U. T. 1956. Is Consciousness a Brain Process? *British Journal of Psychology* 47: 42–51.

Polger, Thomas. 2003. *Natural Minds*. Cambridge, MA: MIT Press.

Primas, H. 1983. *Chemistry, Quantum Mechanics, and Reductionism*. Berlin: Springer-Verlag.

—— . 1991. Reductionism: Palaver Without Precedent. In *The Problem of Reductionism in Science*, edited by E. Agazzi. Dordrecht: Kluwer.

—— . 1998. Emergence in the Exact Sciences. *Acta Polytechnica Scandinavica* 91: 83–98.

Prinz, Jesse. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Putnam, Hilary. 1961/1980. Brains and Behavior. In *Readings in the Philosophy of Psychology*, edited by N. Block. Cambridge, MA: Harvard University Press. (Originally published in *Program of American Association for the Advancement of Science, Section L [History and Philosophy of Science]*, December 27.)

—— . 1975. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131–93.Pylyshyn, Zenon D. 1999. Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception. *Behavioral and Brain Sciences* 22 (3): 341–423.

Quillian, M. R. 1968. Semantic Memory. In *Semantic Information Processing*, edited by M. Minsky. Cambridge, MA: MIT Press.

Quine, W. V. O. 1953. *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

—— . 1960. *Word and Object*. Cambridge, MA: MIT Press.

—— . 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramsey, William, and Stephen P. Stich. 1990. Connectionism and Three Levels of Nativism. *Synthese* 82: 177–205.

Ramsey, William, Stephen P. Stich, and J. Garon. 1991. Connectionism, Eliminativism, and the Future of Folk Psychology. In *Philosophy and Connectionist Theory*, edited by W. Ramsey, S. P. Stich, and D. Rumelhart. Hillside, NJ: Erlbaum.

Robinson, William. 2004. *Understanding Phenomenal Conscousness*. Cambridge: Cambridge University Press.

Rorty, Richard. 1979. *Philosophy and the Mirror of Nature*. Oxford: Blackwell.

Rosenberg, Alexander. 1994. *Instrumental Biology or the Disunity of Science*. Chicago: University of Chicago Press.

Rosenthal, David. 1986. Two Concepts of Consciousness. *Philosophical Studies* 49: 329–59.

Rouse, Joseph. 1987. *Knowledge and Power*. Ithaca, NY: Cornell University Press.

—— . 1998. New Philosophies of Science in North America: Twenty Years Later. *Journal for the General Philosophy of Science* 29 (1): 71–122.

—— . 2003. *How Scientific Practices Matter: Reclaiming Philosophical Naturalism*. Chicago: University of Chicago Press.

Roush, Wade. 2006. Marvin Minsky on Common Sense and Computers That Emote. *Technology Review*, July 13. Available at: http://www.technologyreview.com/read_article.aspx?id=17164&ch=infotech)

Sacks, Oliver. 1985. *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. New York: Simon and Schuster.

Salmon, Wesley. 1971. *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

—— . 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Sarkar, Sahotra. 1992. Models of Reduction and Categories of Reductionism. *Synthese* 91: 167–94.

Sayre, Kenneth. 1986. Intentionality and Information Processing: An Alternative Model for Cognitive Science. *Behavioral and Brain Sciences* 1 (March): 121–38.

Schaffner, K. F. 1967. Approaches to Reduction. *Philosophy of Science* 34: 137–47.

—— . 1974. The Peripherality of Reductionism in the Development of Molecular Biology. *Journal of the History of Biology* 7: 111–39.

Schank, R. C., and R. P. Abelson. 1977. *Scripts, Goals and Understanding*. Hillside, NJ: Erlbaum.

Schliesser, Eric. 2004. The Missing Shade of Blue, Reconsidered from a Newtonian Perspective. *Journal of Scottish Philosophy* 2: 164–75.

Searle, John. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences* 3: 415–57.

—— . 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sellars, Wilfrid. 1960. Being and Being Known. *Proceedings and Addresses of the American Philosophical Association* 34 (1660): 28–49.

Shapere, Dudley. 1984. *Reason and the Search for Knowledge*. Dordrecht: D. Reidel.

Sharma, A., A. Angelucci, and M. Sur. 2000. Induction of Visual Orientation Modules in Auditory Cortex. *Nature* 404 (6780): 820–21.

Siewert, Charles. 1998. *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.

Silberstein, Michael. 2002. Reduction, Emergence and Explanation. In *The Blackwell Guide to the Philosophy of Science*, edited by P. Machamer and M. Silberstein. Malden, MA: Blackwell.

Simon, Herbert. 1977. *Models of Discovery and Other Topics in the Methods of Science*. Dordrecht: Reidel.

Sklar, Lawrence. 1999. The Reduction (?) of Thermodynamics to Statistical Mechanics. *Philosophical Studies* 95 (1–2): 187–99.

Smart, J. J. C. 1959. Sensations and Brain Processes. *Philosophical Review* 68: 141–56.

Spelke, Elizabeth. 2002. Developing Knowledge of Space: Core Systems and New Combinations. In *Language of the Brain*, edited by S. Kosslyn and A. Galaburda. Cambridge, MA: Harvard University Press.

——— . 2003. What Makes Us Smart? Core Knowledge and Natural Language. In *Language in Mind*, edited by D. G.-M. Gentner. Cambridge, MA: MIT Press.

Sperber, D. 1994. The Modularity of Thought and the Epidemiology of Representations. In *Mapping the Mind*, edited by I. Hirschfeld and S. Gelman. New York: Cambridge University Press.

Sterelny, Kim. 2003. *Thought in a Hostile World*. Oxford: Blackwell.

Stich, Stephen P. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Stich, Stephen P., and Stephen Laurence. 1994. Intentionality and Naturalism. In *Midwest Studies in Philosophy*, edited by P. A. French, T. E. Uehling, Jr., and H. K. Wettstein. Notre Dame, IN: University of Notre Dame Press.

Strawson, Peter F. 1959. *Individuals*. London: Methuen.

Suppes, Patrick. 1974. The Structure of Theories and the Analysis of Data. In *The Structure of Scientific Theories*, edited by F. Suppe. Urbana: University of Illinois Press.

Sur, M., P. E. Garraghty, and A. W. Roe. 1988. Experimentally Induced Visual Projections into Auditory Thalamus and Cortex. *Science* 242 (4884): 1437–41.

Sutton, W. S. 1903. The Chromosomes in Heredity. *Biological Bulletin* 4: 24–39.

Tarski, Alfred. 1944. The Semantic Conception of Truth and the Foundations of Semantics. *Philosophy and Phenomenological Research* 4: 341–75.

Thagard, Paul. 1988. *Computational Philosophy of Science*. Cambridge, MA: MIT Press.

——— . 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Toulmin, Stephen. 1972. *Human Understaning*. Vol. 1. Princeton, NJ: Princeton University Press.

——— . 1974. The Structure of Scientific Theories. In *The Structure of Scientific Theories*, edited by F. Suppe. Urbana: University of Illinois Press.

Tritton, D. J. 1976. *Physical Fluid Dynamics*. Oxford: Oxford University Press.

Trumpler, M. 1997. Converging Images: Techniques of Intervention and Forms of Representation of Sodium-Channel Proteins in Nerve Cell Membranes. *Journal of the History of Biology* 30: 55–89.

van Fraassen, Bas. 1980. *The Scientific Image*. Oxford: Clarendon.

——— . 1988. The Pragmatic Theory of Explanation. In *Theories of Explanation*, edited by J. Pitt. New York: Oxford University Press.

van Inwagen, Peter. 1993. *Metaphysics*. Boulder, CO: Westview Press.

Wagner, Steven J., and Richard Warner. 1993. *Naturalism: A Critical Appraisal*. South Bend, IN: University of Notre Dame Press.

Wilson, E. O. 1998. *Conciliance: The Unity of Knowledge*. New York: Knopf.

Wilson, Mark. 2006. *Wandering Significance*. New York: Oxford University Press.
Wimsatt, William C. 1974. Complexity and Organization. *PSA* 1972: 67–82.
——— . 1976a. Reductionism, Levels of Organization, and the Mind-Body Problem. In *Consciousness and the Brain: A Scientific and Philosophical Inquiry*, edited by G. G. Globus, G. Maxwell, and I. Savodnik. London: Plenum Press.
——— . 1976b. Reductive Explanation: A Functional Account. *PSA* 1974: 671–710.
——— . 1980a. Randomness and Perceived-Randomness in Evolutionary Biology. *Synthese* 43: 287–329.
——— . 1980b. Reductionistic Research Strategies and Their Biases in the Units of Selection Controversy. In *Scientific Discovery: Case Studies*, edited by T. Nickles. Dordrecht: Reidel.
Winograd, Terry. 1972. *Understanding Natural Language*. Orlando, FL: Academic Press.
Winograd, Terry, and Fernando Flores. 1986. *Understanding Computers and Cognition*. Norwood, NJ: Ablex.
Winsberg, Eric. Forthcoming. Handshaking Your Way to the Top: Inconsistency and Falsification in Intertheoretic Reduction. *Philosophy of Science*.
Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media.
Wurtz, R. H., and E. Kandel. 2000. Perception of Depth, Motion and Form. In *Principles of Neural Sciences*, edited by E. Kandel, J. H. Schwartz, and T. M. Jessell. New York: McGraw-Hill.

This page intentionally left blank

Index