A Place for Consciousness: Probing the Deep Structure of the Natural World

Gregg Rosenberg

OXFORD UNIVERSITY PRESS

A Place for Consciousness

PHILOSOPHY OF MIND SERIES

Series Editor
David J. Chalmers, University of Arizona

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

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

Three Faces of Desire
Timothy Schroeder

Gut Reactions
A Perceptual Theory of Emotion
Jesse J. Prinz

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

A Place for Consciousness

Probing the Deep Structure of the Natural World

Gregg Rosenberg

OXFORD
UNIVERSITY PRESS
2004

OXFORD
UNIVERSITY PRESS

Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Taipei Tokyo Toronto

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

www.oup.usa.org

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Rosenberg, Gregg.
A place for consciousness : probing the deep structure of the natural world / Gregg Rosenberg.
p. cm.—(Philosophy of mind series)
Includes bibliographical references and index.
ISBN 0-19-516814-3
1. Consciousness 2. Philosophy of nature. I. Title. II. Series.
B808.9.R67 2004
126—dc22 2003063988

1 3 5 7 9 10 8 6 4 2

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

Do you believe that absolutely everything can be expressed scientifically?

—Hedwig Born to Albert Einstein

Yes, it would be possible, but it would make no sense. It would be description without meaning, as if you described a Beethoven symphony as a variation of wave pressure.

—Einstein's reply

This page intentionally left blank

This book is dedicated to the memories of my father,
Donald Rosenberg, and my good friend David Han.
I loved you both. Rest in peace.

This page intentionally left blank

Preface

My intention in writing this book was to create something whose importance lies beyond the details of its arguments. I myself consider this primarily a book of ideas. Of all my hopes, my dearest is this: that *A Place for Consciousness* should provide inspiration to those like me who were raised with the physicalist orthodoxy, accepting it but not fully comfortably, whose disquiet always has been silenced at the end by the baffling question: *How could it be otherwise*? I believe this book points to a place in the space of philosophical ideas where something truly new and interesting exists. I am, above all, trying to lead readers to that place so that they can return without me to explore it on their own. The space of ideas is a public space, after all, and these particular hidden woods can surely be mapped better than I have been able to map them.

We all know that in some sense there *is* a ghost in the machine. The question that grips us is, *why*? Why does consciousness even exist? What use has nature for an experience machine? This book proposes a place for consciousness in nature. The framework developed here is ambitious in its scope and detail: It ties experience into a theory of the categorical foundations of causation. Scholars should see it as an attempt to make a substantial advance in the development of Bertrand Russell's *Structural Realism* by borrowing some inspiration from Alfred North Whitehead's process philosophy. General readers can simply see it as an attempt to explain the mystery of the soul. *Liberal Naturalism* is my name for views of this type.

Both Russell and Whitehead argued that physical science reveals only a structural aspect to nature. If physics is all structure, it is natural to suppose that intrinsic properties related to the intrinsic properties we experience in consciousness are the intrinsic content of the physical. This suggestion raises several questions: (1) Why should the intrinsic properties of a physical system be experiential? (2) Why do they exist above the level of the microphysical, where large-scale cognitive systems might experience macrolevel intrinsic content? (3) Why should they form a unity of the kind we are acquainted with in consciousness? and (4) Why should phenomenal content, as the intrinsic content of the physical, correspond so closely to the information structure within the brain? By constitutively linking experience and causation, I answer these questions from first principles.

This may seem like an unlikely project because the two problems of consciousness and causation are each tough philosophical chestnuts individually. It is not clear that thumping them together will really help us crack them open. I hope to meet the burden of the project: to argue that they need to be treated together and to show, in a very concrete way, how they do go together. To meet my obligations, I argue that physicalism is false, yet I also show how one can reject physicalism in a way that is perfectly compatible with physical science. This is a tough ledge to walk. Accordingly, the aims I have for this work extend only to motivating, introducing, explaining, and defending the overall framework, while leaving detailed discussion of its applications to a sequel. I divide my aims into several levels of ambition even within these boundaries.

At the first level of ambition, I wish to *provoke*. Within the book, I defend a group of ideas that are at odds with the physicalist orthodoxy within science and the philosophy of mind. I believe the framework I flesh out here should at least make physicalists uncomfortable by showing that a nonphysicalist theory need not be supernatural, naturalistically untenable, unmotivated, or hopelessly vague. After reading it, no one should rest comfortably with any assumption that alternative views to physicalism must lead to absurdity.

At the next level of ambition, I hope to *challenge*. Physicalism's strongest support has been the widespread intuition that only physicalism can guarantee the causal relevance of experience in an acceptable way. A first challenge coming out of this book is that, by explaining why physics is not a theory of causation, it is able to show vividly why the issue makes sense only against a detailed background theory of causation. We see, furthermore, that traditional fears about alternatives to physicalism are without support under at least one possible and substantial view of causation, a view that seems compatible with physical science. Not only does experience turn out to have a place in the causal order on the Liberal Naturalist view, but I also make a case *on grounds completely independent of the mind-body problem* that something exactly like it, in its most mysterious aspects, is required for causation to exist.

A second challenge, one for those sympathetic with the project begun in this book, is to see whether the ideas here lead to fruitful avenues of research or whether, instead, they lead down a dead end. The book only presents a framework called the Theory of Natural Individuals. This framework should provide a new perspective from which to understand nature and many open questions about applying the framework remain at the end of this work. These open questions present the possibility for an actual empirical and philosophical research program. It is particularly important to discover the details about the physical conditions that correspond to the existence of the things I call *natural individuals* in the book.

At a third level, I hope to actually *convince*. Although I propose some unusual ideas here, I take no shortcuts, and I accompany my proposals with substantive discussion and argument. *Liberal Naturalism* is currently a minority position, but it at least has current precedents within philosophy, especially in the work

of philosophers such as David Chalmers, David Griffin, Daniel Stoljar, Galen Strawson, and Michael Lockwood.

My more specific proposal, which I call *the Theory of Natural Individuals*, involves experience directly in the fundamental causal character of the world. This more specific proposal seems very radical when stated baldly, but I have not pulled a rabbit out of a hat: Nowhere in this book will the reader find a conjuring trick, a ploy of misdirection, or a wave of the hands. I have tried to work with acceptable rigor by generalizing on some fairly mundane intuitions about the world and about consciousness. And I have tried, always, to respect science. I hope that I have succeeded in rationally motivating my case and that the work is potentially fruitful.

As a work of philosophical literature, *A Place for Consciousness* began in 1988 while I was pursuing my master's degree in Artificial Intelligence. I worked rather doggedly at trying to map the terrain for nearly ten years, resulting in a too-rough first attempt at putting it all together in my 1997 dissertation in philosophy and cognitive science. The year before that, David Chalmers released his book *The Conscious Mind*. As I set about trying to tame the wild threads of my dissertation work into something mature and more polished, I initially conceived of this book as a kind of unauthorized sequel to David's book. In time, I realized that he had set the bar too high for me. I hope instead to have produced at least worthwhile companion reading.

While this book is by no means an easy read, I have aimed to make it accessible and interesting to the generally educated and intellectual public, even to those who have little or no training specifically in philosophy (with the exceptions of chapters 3 and 10, which are necessarily technical). Although the book is long, it is possible to take a short tour and still come away with the main ideas. For those interested in the short tour, I recommend reading chapters 1 and 2 to understand the setup of the problem. From there, skip to chapters 4, 9, and 12. If the short tour piques your interest, go back and read the rest. Those with a philosophical background who are comfortable with one or more of the standard responses to the antiphysicalist arguments should read chapter 3. Also, the remaining chapters in Part I provide more thorough reasons than the short tour does for believing that someone interested in understanding consciousness should look hard at causation itself. Finally, Part II may be interesting independently of one's views on the mind-body problem, especially the arguments against Humean views in chapter 8 and the detailed treatment of the causal nexus in chapters 9 through 11.

This page intentionally left blank

Acknowledgments

Reinventing nature is hard work. I could not have done even the little bit of it that I do here without a lot of support from others, both intellectually and emotionally. In my lifetime, my interests have taken me down many paths. Each part of me has found some reflection in this book, and I am indebted to many who helped to steer me down my long and winding road.

I thank Anthony Nemetz for first introducing me to the world of intellectual questioning when I was an undergraduate business major. His demanding eloquence was a revelation to me at that time in my life, as nothing in my background had previously exposed me to intellectual life.

I owe my deepest debts from my time at the University of Georgia to Donald Nute. Not only did he direct my master's thesis when I was studying Artificial Intelligence there, but he has encouraged and supported me every step of the way since: first in my decision to move into philosophy, then by encouraging me to go to Indiana University to do my Ph.D., and finally by accepting me back at the Artificial Intelligence Center as a postdoctoral researcher.

I thank Ned Block for the helpful conversations we had during my time at MIT in 1991. His insistence that ideas as unusual as mine need to be very strongly motivated has always stuck with me, acting as a burr whenever I have been tempted to cut corners in my writing or thinking.

I thank Douglas Hofstadter, whose books *Metamagical Themas* and *Gödel, Escher, and Bach* serendipitously fell into my hands while I was an undergraduate, steering me toward the philosophy of mind and cognitive science. My eventual interactions with him while pursuing my Ph.D. at Indiana University were challenging and provocative.

When I began my graduate work at Indiana University, I came to school convinced about the explanatory gap between the facts of consciousness and the physical facts, and I suspected that there must be a deep link between consciousness and causation itself. I was extremely fortunate to arrive there at the same time at which David Chalmers was finishing his dissertation on consciousness in the same program. I have since discovered that the only thing comparable to David's intellect is his generosity of spirit. First, I thank David for clearing a path that has made a book such as this one possible. Second and most importantly, I

thank David for his friendship, for our many conversations and correspondences, and for his continued assurance that this work is interesting and worthwhile. Finally, I thank him for especially helpful comments on how to best organize the material in chapters 2 and 3. If I had always listened, then I am sure those chapters would be better.

These ideas were first written in preliminary form as my dissertation in Indiana University's Philosophy and Cognitive Science program. Mike Dunn chaired my dissertation committee and gave generously of his time and advice. Our conversations ranged freely around the philosophical world, from topics such as Platonism to the nature of properties to the nature of implication to the nature of mind. His restraint in passing harsh judgment on my speculations, choosing instead to ensure that I asked myself the right questions, made me feel that I had a right to travel over the wide terrain I cover in this book. I thank him for providing his comments and support at such a crucial time.

I give special thanks to Anil Gupta, not only for the helpful discussions we have had over the years but also for providing me with a role model for the way a true philosopher should conduct himself. His probity, patience, gentleness, and integrity have been an inspiration to me. I thank Tim O'Connor for his enthusiasm, incredible energy, and time at our long lunches. His ideas on how to do metaphysics seriously have been invaluable. In my last year at Indiana, I was very fortunate to meet Brian Cantwell Smith. Like me, Brian is a computer scientist-cum-philosopher, and the perspective that gives is difficult to put into words. I am grateful to Brian for the long hours he gave trying to help me improve my writing.

I also thank John Gregg for supportive encouragement and extremely helpful feedback on drafts of this book. John is owed a special round of thanks because the effort he put into commenting on a draft of this book chapter by chapter resulted in some substantial improvements in clarity. William Seager and Torin Alter also took the time to read the entire work in manuscript form and provided much-needed feedback and support. I also owe thank-you's to Brie Gertler and Brad Thompson for helpful comments on parts of chapter 3.

Nothing can substitute for heated arguments over beer that last late into the night. I have almost too many of these informal debts to list, mostly to my fellow graduate students while I was in the philosophy department at Indiana University. I would like to single out for special thanks a handful who have provided especially memorable philosophical conversation: Tony Chemero, Diarmuid Crowley, Stephen Crowley, Eric Dalton, Craig DeLancey, Jim Hardy, Anand Rangarajan, and Adam Kovach.

I owe my warmest thank-you's to Leslie Gabriele. Not only has she provided me with an important intellectual sounding board, but also her friendship and support were priceless on a personal level. I would not have gotten through some of the rougher times over the past few years without her.

Along those same lines, I would like to thank my long-time friends, especially Allen Domenico, Scott Davis, and Bob Lauth, for their support and encourage-

ment. The most precious friendships are the ones that you know will last a lifetime.

My deepest thanks are reserved for my mother, Sally, my late father, Donald, and my brother, Alan. They have made an investment in my life and identity that is truly staggering to consider. Every word in here reflects their love.

The bulk of the writing of this book occurred in three bursts, enabled by support from outside sources. In 1996–1997, I first formulated the basic ideas expressed here as my dissertation, and I could not have done nearly what I did without dissertation-year support from the Nelson Foundation and the Institute for Humane Studies. In 1998–1999, I was able to advance the ideas in my dissertation and produce a first draft of the book while a Fetzer Fellow, and I thank the Fetzer Foundation for their confidence in my work. I thank the University of Georgia's Artificial Intelligence Center for providing a supportive environment while I was a Fetzer Fellow. The book was stabilized and made ready for publication in 2001, after I had the good fortune of selling my Internet security company, and I am grateful for the free time I have had since.

Finally, none of the people to whom I owe these debts are responsible for any errors in fact, scholarship, judgment, omission, or organization in this book. I claim its shortcomings all for my own.

This page intentionally left blank

Contents

PART I

Liberal Naturalism

This page intentionally left blank

1

A Place for Consciousness

1.1 The Topic

Consciousness is a refugee. It gathers the interest and sympathy of many disciplines without claiming a true home in any of them. Often abused by skeptics, it has been exploited by dreamers. Until recently it was ignored by experimentalists, and theorists have not always taken it seriously. If any important piece of nature could lay claim to being an intellectual exile, consciousness has been it. The purpose of this book is to find a place for consciousness.

Consciousness is an ambiguous term[1] and not all senses of the term pose the same kinds of problems. The central problem it poses is where in nature to place *subjective experience*, which is responsible for the subjective quality of our existence. Philosophers call this sense of consciousness *phenomenal* consciousness. Phenomenal consciousness is special. It is different from just wakefulness, for instance. Dreaming is a way of experiencing, and, therefore, in the sense that needs placement, we are conscious during sleep.

Phenomenal consciousness is not necessarily consciousness *of* anything *else*. For example, when I close my eyes and cover my eyelids with the palms of my hands, I see diffuse shapes floating in the blackness and jumpy patches of diluted color. These are experiences and are thus elements of phenomenal consciousness, even though they do not seem to *represent* anything.

Phenomenal consciousness does not necessarily involve language or self-understanding. For example, when a newborn infant cries on first experiencing the world, it must be *feeling* something, even though it has not yet developed language or self-understanding. Because it feels, it is phenomenally conscious.

We identify phenomenal consciousness by being acquainted with it, not by looking up a scientific definition. Even though "phenomenal consciousness" does not have a scientific definition yet, I mean phenomenal consciousness when I use the word *consciousness* in this book. If we need a definition, the best we can do

is to create an operational definition by calling attention to it in increasing levels of detail.

The most succinct way to convey the meaning of the term is through Thomas Nagel's popular phrasing from 1974: A creature's subjective experience constitutes *what it is like* to be that creature. For example, part of what it is like to be a person with normal color vision is for purple things to subjectively appear in a certain way, as having a certain kind of visual quality to that person. Purple subjectively appears different from pink, which is subjectively different from orange, which is subjectively different from black, and so on. Together, the subjective appearances of these qualities help make up *what it is like* to be a person with normal color vision.

After becoming aware of these visual qualities *as* qualities, you may naturally wonder what the colors from a larger color space *look like*. For example, some birds can see colors that no person can see. What is the experience like when these birds see the extra colors available to them? Once you know about their ability, a question about the character of their conscious experience remains. The facts about these birds' phenomenal consciousness include what it is like for them to see the extra colors they see.

Similarly, just as the subjective qualities involved in seeing something (e.g., colors, shape, and depth) are different in kind from the ones involved in hearing something (e.g., tone, pitch, and rhythm), there must be a set of distinct qualities that make up what it is like for a bat using its echolocation. Are the qualities that the bat experiences like those you experience when seeing something, or are they like those you experience when hearing something, or are they like something else altogether? In the same spirit, you may also wonder what the qualities and sensations associated with a manta ray's sensing of electromagnetic currents on the ocean floor are like for the manta ray.

Examples multiply easily. Philosophers call the subjective qualities these questions point to *phenomenal qualities*, or *qualia*. At the extreme, you may even wonder, however implausibly, whether it is like *anything at all* to be these creatures. Perhaps they are unconscious robots, all "dark inside," without any qualia at all.[2]

Phenomenal consciousness is richly varied, complex, and subtle. For example, the exact organization of the qualities of experience, and perhaps even their character, seems to be very responsive to conceptualization. An example of this occurs when we stare at visually ambiguous figures such as the Necker cube in figure 1.1: The qualitative experiences associated with seeing its face *as* oriented upward or *as* oriented downward are very distinct. This suggests a location for the world's repository of facts concerning phenomenal consciousness. For a particular creature, the facts concerning what it is like to be that creature are constituted by (1) its capacities for experiencing phenomenal qualities in the first person and (2) its way of conceptualizing the world.[3]

What is the place of consciousness in our world? From where does phenomenal information come? Are phenomenal facts ordinary physical facts? Are they

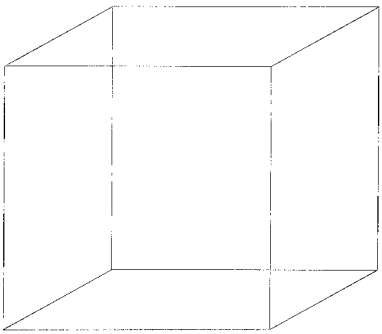


Figure 1.1 A Necker cube. When we stare at the Necker cube, our phenomenal experience changes depending on whether we perceive it as facing upward or downward.

the kinds of facts that ordinary physical facts can form a basis for? And, if so, in what way can physical facts provide a basis for them? We do not have good answers to these questions yet.

Moving just slightly beyond Nagel's slogan, Brian Loar (1990) delivers a longer description of the intended target by concisely expanding the slogan, What it is like to be*:*

> On a natural view of ourselves, we introspectively discriminate our own experiences and thereby form conceptions of their qualities, both salient and subtle. These discriminations are of various degrees of generality, from small differences in tactual color experience to broad differences of sensory modality, e.g. those among smell, hearing and pain. What we apparently discern are ways experiences differ and resemble each other in respect of what it is like to have them. Following common usage, I will call those experiential resemblances phenomenal qualities; and the conceptions we have of them, phenomenal concepts. Phenomenal concepts are formed "from one's own case". They are type-demonstratives that derive their reference from a first-person perspective: "that type of sensation", "that feature of visual experience". And so third-person ascriptions of phenomenal qualities are projective ascriptions of what one has grasped in one's own case: "she has an experience of that type."

I want to clarify Loar's characterization in one important respect. Rather than using *phenomenal qualities* to denote resemblance between experiences, I use the phrase to denote the qualities within experience that are *responsible for* these resemblances between them.

At the next level of detail, you can catalogue varieties of phenomenal experience by paying close attention to the different kinds of experiences you can have. Cataloguing exercises can direct and refine your awareness of the subject matter by highlighting for you your own subjective acquaintance with the characters of your inner life. David Chalmers catalogues experience in the first chapter of his *The Conscious Mind* (1996). He calls attention to, and gives short accounts of, the fascinating variety of phenomenal content found in experiences as diverse

as: visual, auditory, tactile, olfactory, and taste experiences; experiences of temperature; pains; other kinematic and proprioceptive sensations; mental imagery; conscious thought; emotions; and the sense of self. When thoughtfully done, catalogues vividly create awareness of phenomenal consciousness and its many elements and forms.

At the most extreme level of detail, you can isolate the meaning of *phenomenal consciousness* by comparing and contrasting it with other senses of the term *consciousness*. Ned Block (1995) does this in a concise way by comparing and contrasting "consciousness" in the sense of having cognitive access to information with "consciousness" as experience. Charles Siewert's (1998) *The Significance of Consciousness* contains an extremely detailed attempt to isolate the sense of the term that picks out the mystery, drawing it out from its hiding place among the other senses of the term.

1.2 The Mind-Body Problem

If you want to understand the problem, Descartes is a good place to start. René Descartes is often credited with creating the modern form of the question, What is the relationship between the mind and the body? This is the *mind-body problem.*

Descartes believed in a metaphysics of substance and properties. A *substance* is supposed to be the metaphysical substrate that supports the existence of properties. *Properties* are repeatable characteristics of things, in the sense that many different things can have the same property. For instance, mass is a property, as many different things can have mass.

Descartes proposed that the substance *matter* essentially has properties of spatial extension and causal power. He also believed that the *mind* is a substance and that it essentially has the properties necessary for rationality and causal power. Beyond this, Descartes believed that rationality was inessential to matter, that spatial extension was inessential to mind, and that, because they have different essential properties, matter and mind could not be the same substance. This is called *substance dualism.*

Substance dualism raises a question about creatures like us who have both minds (composed of the rational substance Descartes called *mind*) and bodies (composed of the spatial substance Descartes called *matter*). How are these substances, which are so different, brought together to be a person?

Descartes suggested that they *interact* with one another through the brain. He admitted to not really understanding how this occurs, but he believed that it must occur. Today we call that position *interactionist dualism*. Together, Descartes's positions made him an *interactionist substance dualist.*

Not many philosophers or scientists today believe in interactionist substance dualism. Most philosophers and scientists believe that mental activity is physically constituted by brain activity. Among academic scientists and philosophers, the most commonly held position is now *physicalism*, which holds that every-

thing is physical *in some sense*. Physicalism is basically the position you would expect to be called *materialism*, except without the historical commitment to the existence of a material substance. In place of Descartes's substances, physicalism just commits itself to the existence of the basic physical properties and events, whatever they turn out to be.

Physicalism belongs to the branch of metaphysics called *ontology*. Ontology is the study of what kinds of things exist, with particular emphasis on the different *ways* of existing possessed by different kinds of things. For example, hurricanes, speed limits, bosons, moral values, numbers, and minds all exist.[4] On their surfaces, at least, these all seem to be very different sorts of things, each with its own unique nature and way of existing.

Ontologists generally focus on two kinds of questions. First, what is the nature of these things? Second, how do all these diverse things come together so that they are able to exist in the same world? Philosophers usually answer this second kind of question by proposing *fundamental* categories of properties, objects, events, or processes whose existence they can see as grounding the existence of other kinds of things. By *fundamental,* philosophers mean that these are the things from which the existence of every other thing is derived.

If one is religious, one may hold that this fundamental thing is God. If one wants a more scientific hypothesis, however, one needs to find another category of things to do this job. That is where physicalism steps in. *Physicalism* is the thesis that all other kinds of things wholly derive their existence from the existence of the physical. Among these *other kinds of things* are hurricanes, speed limits, moral values, numbers, and, most important in this book, conscious minds.

Physicalists often charge Descartes with serious errors that still infect our thinking about the mind. I argue that Descartes's most dangerous errors were the ones he made about matter, not mind. Descartes felt forced to his dualism chiefly because the science of his time had revolutionized our ideas about matter. After the scientific revolution, people thought of matter as something primarily *quantitative* and *geometrical* and best described in terms of how these quantitative states vary at different points in space and time. Thus mathematics and geometry, rather than perception and sensation, came to provide the best models for understanding the essential nature of matter. This revolution in thinking was as radical and important as any intellectual revolution has ever been. I believe it is hard for us now to fully appreciate it.

Prior to this revolution, in which Descartes himself was a leading figure, educated people had primarily thought of matter as something *qualitative*. Qualities are attributes, not necessarily quantitative, found in sensations that make each kind of sensation fundamentally unlike the other kinds. For example, the distinct feelings of itches are qualities and are different from the qualities of smells. Although found in sensations, qualities were thought to exist in matter quite generally, whether sensed or not. Common opinion was that matter is best understood by proposing qualities and investigating how these qualities are *qualified* or *con-*

ditioned through intimate causal relationships that bind them to one another and give them *form*.

This pivotal shift from thinking about matter as something qualitative to thinking about it as something quantitative drew a revolutionary line that has sharply differentiated modern from premodern thinking. In this book, I argue that Descartes's error, and the error that still haunts us, is that we have come to believe that this revolutionary view of matter is *all there is* to matter.

As revealed from a fundamentally Cartesian perspective on the physical, the human body is a marvel whose subtlety, flexibility, and complexity uplift the word *machine*. Natural science tells us that the body is made ultimately of very tiny and exotic physical entities, and we know that it consists in the motions of, and interactions among, delicately layered physical structures. Our bodies are spatiotemporal organizations of these tiny entities, driven by an enormous number of microphysical interactions.

From this perspective, the mind-body problem arises immediately: How could a collection, *any* collection, of microphysical interactions have macrolevel *experiences*? According to physical theory, the entire being of these microphysical entities consists in the quantitative dispositions that produce their intrinsic dynamics and their intimate couplings. The mystery of consciousness is the question of why this assembly, this whirlwind of causation, should ever *feel*. Couldn't this causation go on without feeling, without sensation, without experiencing at all? Viewed in the large, these finely layered patterns are dynamical wonders, but it is hard not to wonder why the dynamics should be conscious. Physical causation produces changes in quantity, shape, and motion, but why should a congeries of quantity in motion, however complexly shaped, ever experience the delightful sweetness of cheesecake? Questions such as this pose the greatest obstacle to the challenge of naturalizing the mind.

1.3 Liberal Naturalism

Even though I argue against physicalism, I am a *naturalist*. The view I favor is Liberal Naturalism. I view naturalism as a methodological requirement to place human beings in the world without making special, ad hoc assumptions that are discontinuous with everything else we have good reason to believe about nature. A fundamental message of this book is that we have good reasons, reasons independent of mind, to understand nature differently than physicalists typically do, and I propose a specific way of doing it that allows us to find a place for consciousness.

The position I develop is a kind of dual-aspect view that I think respects what is right about the intuitions of both physicalists and substance dualists. Dual-aspect views provide an alternative to substance dualisms for antiphysicalists. Whereas substance dualism proposes that there are two fundamentally different and potentially independent kinds of entities, matter and mind, dual-aspect views hold that there is one fundamental kind of entity but that this entity has more than a physical aspect. It is like the difference between thinking the evening star

and the morning star are different stars and thinking that they present different aspects of the same thing, the planet Venus.

Like physicalism, Liberal Naturalism holds that the world is probably composed from a single fundamental kind of thing. This fundamental kind of thing, if it exists, probably has a set of fundamental properties that are mutually related in a coherent and natural way by a single set of fundamental laws. However, like substance dualism, Liberal Naturalism holds that some of these properties and laws are not physical properties and laws. What ties the physical and nonphysical together is a deeper kind of thing of which they are both aspects.

As a Liberal Naturalist, I identify (to a greater or lesser degree) with David Chalmers, Thomas Nagel in some of his moods (e.g., his 1998 work), Wilfrid Sellars on some ways of reading his work (e.g., his distinction between his physicalism$_1$ and physicalism$_2$), Abner Shimony, Grover Maxwell in his writings on *structural realism* (1971, 1979), Michael Lockwood, Alfred North Whitehead, David Ray Griffin, and Bertrand Russell in his neutral monist phase. The Liberal Naturalists recognize the possibility that the specifications of physics and what could subsist in a world wholly portrayed by physics may not circumscribe nature's limits. That allows the Liberal Naturalist to step comfortably outside the standard physicalist ontology while retaining a naturalist outlook.

The positive project in this book is to identify what these nonphysical properties are; to explain why they should exist; and to give reasons for believing they fit cohesively within a scientific and naturalistic worldview. I pursue these goals by introducing a substantive view of causation. This metaphysically rich picture of causation provides the bridge that takes us from the physical to consciousness. It also respects the causal closure of the physical, as I attempt to complete our view of causation by adding elements that are complementary to the structure of activity described by physical science and that, for that reason, are every bit as essential to it as is the physical.

I make and develop several distinctions between aspects of causation, including:

1. Distinguishing the *effective* properties as properties that give individuals the inherent potential to place constraints on one another.
2. Developing a theory of shared *receptivity* to provide a context in which the effective properties can be realized and do their work, thus forming the basis of the *connectivity* between individuals.
3. Proposing that the effective and receptive causal dispositions must be *carried* by fundamental intrinsic properties. It is through understanding these carriers that we can understand why consciousness exists.

After developing this model, I argue that physics describes only spatiotemporal patterns in the appearances and values of effective properties. I argue that a realist account of the causal nexus goes beyond this physical aspect because physical theory leaves out information about receptive connectivity and the intrinsic carriers. It follows that a complete theory of the causal nexus needs to go beyond physical theory.

If the model I propose and develop in this book is right, experiencing is a fundamental element of nature. It has a natural place in the implementation of causation, and phenomenal qualities implement nature's effective constraints. In the terminology I introduce later, experiencing acts as an intrinsic carrier for causation itself. The phenomenal qualities carry the effective properties of individuals within a causal nexus, and the experiencing of these qualities carries the receptiveness had by members of the nexus to these effective properties.

It turns out that the place of consciousness in the natural world intrinsically connects it to a larger, metaphysical background via its intimacy with causation itself. Under the kind of realist account of causation I detail, a picture emerges that does not drive a wedge between consciousness and the physical world. Instead, it locates us within a world that is richer both naturally and metaphysically than the one previously available. The resulting view avoids the interaction of Descartes's substance dualism without slipping into the brute and inexplicable identities of physicalism, and so provides the foundations for a possible Liberal Naturalism.

1.4 The Structure of the Book

The main body of the book is divided into two parts. In part I, "Liberal Naturalism," I first argue that physicalism cannot adequately account for consciousness. To establish physicalism's failure, I analyze what it means to be a physical fact by establishing an analogy with an artificial kind of world. The analysis shows, in a concrete way, why no physicalistic theory will entail the facts of consciousness and defends the importance of entailment to the truth of physicalism. The failure of physicalism creates a puzzle regarding just what consciousness might be, if not physical.

After presenting this puzzle, I explore problems and tensions created by the implication that there must exist fundamental nonphysical properties. *How* can the world have both physical and phenomenal aspects? And *why* would it? By searching the places at which these two aspects seem most incompatible with each other, I try to discover clues about where the incompleteness in our knowledge might lie. Among other conclusions, I argue that the existence of consciousness is evidence for hidden structure within nature. Also, I argue that, at every turn, our search points us toward the need to more fully understand causation itself.

Perhaps the metaphysics of causation is richer than materialists usually suppose. I devote part II of the book, "Faces of Causation," to a direct analysis of causation and the conditions on the possibility of causal interaction. As a first point, I build a case that the explanation of causation also requires nature to have multiple aspects: its effective aspect, its intrinsic connectivity, and the intrinsic carriers of the causal dispositions. I build a speculative metaphysics for causation, a metaphysics in which the roles of each type of element are specified in a rigorous way. The detail of my development allows me to place consciousness in

the world in a way that answers the puzzles, paradoxes, and tensions I raise in part I while avoiding the usual objections to dualist views.

1.5 The Sliding Tile Puzzle

The mystery of consciousness is both profound and exciting. If one thinks hard and long about it, the questions it raises will linger and endlessly deepen. Eventually, they seem to transcend specific questions about consciousness, touching insecurities about our understanding of nature herself. At first one tries to solve the puzzle as though it were a jigsaw puzzle, with pieces nestled stably in their proper places. Eventually one begins to realize that, to solve the puzzle of conscious experience, we may have to view the project as being more like trying to solve a sliding tile puzzle, such as the one in figure 1.2.

A sliding tile puzzle consists of a rectangular frame with movable tiles within it, each tile decorated with a different part of the puzzle. Initially, the tiles are scrambled, and the goal is to unscramble them to retrieve the puzzle's picture. The rectangle contains one empty space, and the puzzle solver must rearrange the

Figure 1.2 Understanding how the pieces of nature fit together is sometimes like trying to solve a sliding tile puzzle.

tiles by sliding them into and out of this empty space. By repeating this, the puzzle solver hopes to undress the confusion and reveal the puzzle's ornamental face.

Sliding tile puzzles contain a trap, a seductive property that lures the unsuspecting. Often the puzzle solver can bring order to *almost* the whole puzzle, perhaps fitting every piece into its proper slot except the last two tiles. These last tiles might be transposed, for instance, each in the other's slots. The trap is sprung when the puzzle solver holds stubbornly to the hard-won order in the rest of the puzzle, afraid that disturbing it too much will cause it to disappear, never to return. Seduced by the order already in the puzzle, the puzzle solver searches desperately for a minimally disruptive solution, one that places the pieces without disturbing the rest of the puzzle very much.

Unfortunately, the puzzle solver cannot usually solve the tile puzzle this way. To fit the final pieces in place, one has to regress first and then rebuild the old order from a new direction. The trap is that, because the puzzle solver flinches at every challenge to the old order, the ideal of completing the puzzle becomes a hopelessly elusive goal. The irony is that the hard-won old order would eventually reappear within a more completely ordered context, if only the puzzle solver could find the strength to first challenge it and, temporarily, relinquish it.

In writing this book, I have approached the problems of consciousness and causation as though they are the final two pieces in a sliding tile puzzle. I wish to help put them into their proper places within a naturalist framework, and I believe that sound arguments exist that this achievement will carry a cost. This cost will require temporarily reneging on some of the hard-won order that science has brought to our understanding of nature. The cost is this: We must concede that physicalism is an inadequate version of naturalism. To justify this cost, I have to touch many other tiles. With luck, the richness of the puzzle will serve to make the effort worth the investment.

2

The Argument against Physicalism

2.1 Introduction

Physicalism says that the fundamental physical facts are the only fundamental facts. All other facts, whether about rocks, tables, morals, or minds, are derivative on these physical facts. In this chapter, I argue that physicalism is false by arguing that a purely physical world could not contain facts of experience. Others have given arguments of this kind, but I hope to look at this kind of argument in a fresh way. In chapter 3 I defend the argument against objections.

My argument is not a form of conceivability argument or knowledge argument. It is a direct argument that the phenomenal facts are of a type that cannot be entailed, either a priori or a posteriori,[1] by the physical facts. To diagnose precisely why entailment fails, I produce a working analysis of physical facts as a type. This working analysis is central to this chapter, and it recurs in part II. Because the specific lessons of this chapter's argument hold recurring importance, I ask even readers who are familiar (or impatient) with the debate over physicalism to pay some attention to this chapter.

2.2 The Dialectic

Recent antiphysicalist arguments rely on thought experiments that claim to show limits on the physicalist program for explanation and, by implication, the metaphysical status of physicalism. In his seminal paper, "What Is It Like to Be a Bat?" (1974), Thomas Nagel argues that any physicalist account of the universe, by being inherently objective, will leave out the subjectivity of points of view. Nagel argues that this omission is reflected in the fact that even when we know all about the physiology of creatures that are very different from us, we do not know what it is like to be them.

Among others, Frank Jackson (1982) and David Chalmers (1996) have refined

Nagel's guiding intuitions. In Jackson's well-known Knowledge Argument, he asks that we consider a superneuroscientist named Mary. From within a black-and-white room, through books and observation of a black-and-white TV, Mary learns everything there is to know about the functioning of the visual system. Jackson maintains that, nevertheless, Mary learns something the first time she is exposed to color. She learns what the experience of blue is like, for instance. Jackson claims that it follows that physicalism must be false because we can know all the physical facts without being able to know, even in principle, *all* the facts.

Chalmers's Conceivability Argument asks us to conceive of a universe physically identical to ours from Big Bang to Big Crunch, but with the twist that our counterparts have no conscious mental life. They are subjective zombies. Chalmers argues that such a universe is conceivable and, furthermore, metaphysically possible. He argues that this shows the falsity of physicalism by showing that the facts about qualitative consciousness are further facts, not determined in the appropriate way by the physical facts.

By using thought experiments, the antiphysicalists aim to show that there is no *entailment* from physical facts to facts about experience, where an entailment is understood as an a priori implication (*A* a priori entails *B* if one can rule out a priori that *A* is true and *B* is false). That is, they aim to show that facts about experience cannot in principle be deduced from physical facts by a priori reasoning. From there, the antiphysicalists argue that physicalism is false. Later I argue against entailment in a different and more general way, using an analogy to an artificial world with a toylike physics. This analogy allows us to diagnose exactly why no kind of entailment, either a priori or a posteriori, can hold in the real world. The result is a direct argument against entailment that does not rely on a conceivability claim or the knowledge argument.

2.3 The Game of Life

Cellular automata names a certain class of artificial, digital worlds. A cellular automaton consists of points, or "cells," located in an abstract space, all of which can have kinds of "causal" properties. Computer modelers define various physics for these worlds and study the behaviors they exhibit. To start an automaton, one assigns an initial distribution of causal properties to the cells, perhaps at random. The automaton then evolves, changing states according to rules that apply pointwise to the space. Typically, the rules that determine which properties a cell will have at a given time are a function on the properties of neighboring cells at an immediately preceding time. One then studies what kinds of entities can evolve and what sorts of properties these entities can have, given the physics that the modeler has created.

Life is the name of a kind of cellular automaton that evolves on a two-dimensional grid. The *Life* world has been used in discussion of the mind-body problem before, most notably in Dennett (1991a), and its physics is extremely simple

and easy to understand. For these reasons, I am also going to use the *Life* world as my example cellular automaton. I define a *pure Life world* as follows:

> Definition 2.1: A world is a *pure Life world* if, and only if, it is a *Life* world of which no fundamental facts are true except those stipulated in its physics.

In *Life*, we are supposed to think of each cell on the grid as a square and as having eight neighbors: a neighbor touching it on each side and a neighbor touching it on each corner. The location of a cell never changes. Additionally, a cell can host exactly one of two mutable causal properties, being *on* or being *off*, at any given time step. To illustrate the basic scheme, figure 2.1 depicts a cell and its neighbors. Three simple rules govern the evolution of a *Life* automaton:

1. If a cell has exactly two *on* neighbors, it maintains its property, *on* or *off*, in the next time step.
2. If a cell has exactly three *on* neighbors, it will be *on* in the next time step.
3. Otherwise, the cell will be *off* in the next time step.

Imagine a *Life* universe consisting of an infinite grid. The two properties possessed by grid cells, *on* and *off*, are the basic physical properties in the *Life* universe. The rules governing the grid's evolution are that universe's laws of physics. When thought about in this way, *Life* becomes a good modeling ground for understanding how physical facts can entail other kinds of facts.

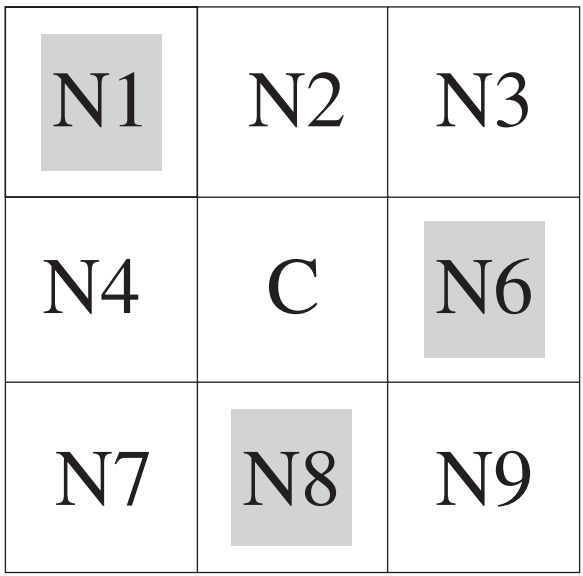


Figure 2.1 A depiction of a center cell, C, and its "neighborhood" on a Life world grid. Neighbor's N1, N6, and N8 are depicted as "on."

Despite its simple physics, the *Life* automaton can evidence a tremendous variety of patterns. For instance, John Conway, the mathematician who invented it, proved that a *Life* grid can be a universal Turing machine.[2] More remarkably, he has proven that the grid can support extremely complex patterns that are self-replicating in von Neumann's sense of nontrivial self-replication (Poundstone 1985). These patterns have functional properties similar to DNA and provide the motivation for the name *Life*. In general, it is the interesting patterns like these in *Life* that create entailments from its basic physical facts to facts of other kinds.

Entities called gliders serve as a simple example of how entailment works in the *Life* universe. A glider consists of a sequence of patterns, each containing exactly five contiguous cells, which move across the grid in a characteristic fashion (see figure 2.2). Gliders make for a useful example because other cellular automata can also produce gliders. This means that *Life* can present *sufficient* conditions for the existence of gliders but cannot present *necessary* conditions, so we cannot *define* the property of being a glider in terms of *Life* physics. To be a glider just means to have a certain structure and to evolve in a certain way, regardless of the underlying physics. The glider structure produces a predictable range of successive states that, lacking interference, move across the grid.

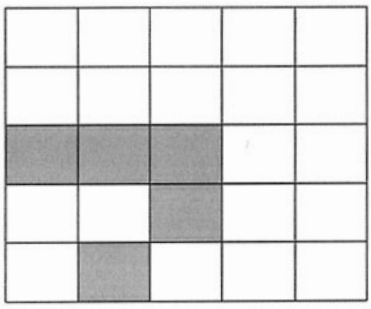
Even before seeing what a stage in the glider pattern actually does when we instantiate it in a *Life* world, we know that *Life* will allow for structure to arise and for the evolution of those structures. Seeing this, it then becomes obvious that *Life* worlds (epistemically) *might* support conditions that entail the existence of gliders. To rule out the (epistemic) possibility that gliders could exist in a *Life* universe, we would need a specific proof that the physics could not produce them.

As it turns out, the *Life* physics can produce gliders. One can prove this by taking a pure *Life* world, producing one of the configurations in the life cycle of a glider, and checking that it evolves correctly over time. It does, so we see that *Life* worlds can entail the existence of gliders.[3]

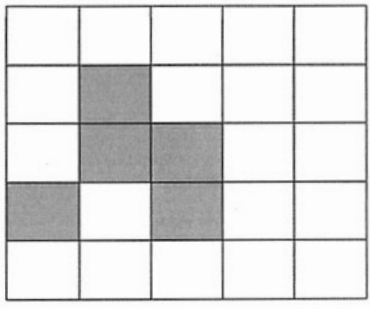
In this example, entailment acts as a determination relation: The basic facts in *Life* are the facts about the distribution of the "on" and "off" properties and how they redistribute over time. Also, the basic *Life* facts necessitate the facts about gliders without our having to introduce any new fundamental ontology. Instead, the necessity is grounded in conceptual truths about what it means to be a glider combined with the empirical truths about the configurations of the basic properties in the *Life* world and the evolution of those configurations. Given a situation in the *Life* world, these interpretive truths are enough to determine the truth of facts about gliders.[4]

2.4 The Form of the Argument against Physicalism

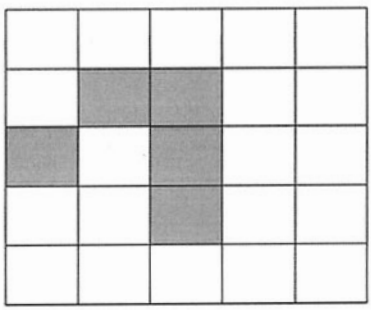
With this understanding of the *Life* world in mind, the argument against physicalism that I defend is:



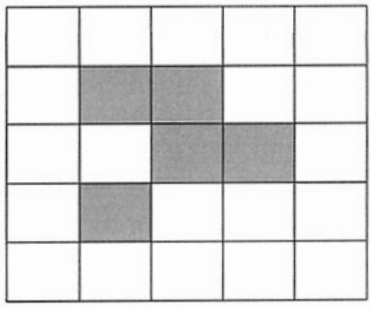
(a) One state of a glider



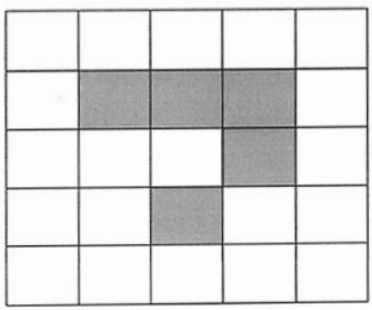
(b) The next state of the glider



(c) The next state of the glider



(d) The next state of the glider



(e) The next state of the glider

Figure 2.2 One full cycle of states in the existence of a glider. Notice that the glider in figure (e) is a copy of the glider in figure (a), only moved up the grid by one cell.

1. Facts about a pure *Life* world do not entail facts about phenomenal consciousness (either a priori or a posteriori).
2. If facts about a pure *Life* world do not entail facts about phenomenal consciousness, then facts about a pure physical world do not entail facts about phenomenal consciousness.
3. Therefore, facts about a pure physical world do not entail facts about phenomenal consciousness.

This is my overall argument. By presenting it, I will lock onto a theoretical conception of what it means to be physical and to be entailed by the physical facts. My strategy is to use the physics of the *Life* world to draw out the categorical structure of physical theories in general, identifying the kinds of information physical theories convey and exposing the kinds of conditions that make physical properties the kinds of properties they are.

2.5 The Argument against Life *Entailing Consciousness*

Facts about a pure Life world do not entail facts about phenomenal consciousness. I defend the first premise of my argument against physicalism by defending something that I call *the Skeptic's Claim*. The Skeptic's Claim is that the facts about a pure *Life* universe cannot entail facts about consciousness. The skeptic's use of "entail" includes both a priori and a posteriori entailment. Thus we may consistently acknowledge any kind of structure and functionality for *Life* objects and still deny the presence of consciousness in a *Life* universe. The argument I defend for the skeptic is:

1. The fundamental properties of a pure *Life* world consist of bare differences.
2. Facts about phenomenal consciousness include facts about qualitative content.
3. Facts about bare difference cannot entail facts about qualitative content.
4. Therefore, some facts about phenomenal consciousness are not entailed by pure *Life* facts.

Premise 1: Pure Life worlds consist of bare differences. What is a *bare difference*? I mean the phrase *bare difference* to express an intuitive idea that can be loosely explained by saying that *Life*'s physics leaves us in the dark about what the "on" and "off" properties are themselves. It just tells us that they are different and enter into certain dynamic relations.

What is an "on" property? It is not the "off" property. What is the "off" property? It is not the "on" property. That, plus the rules of evolution, is all *Life's* physics specifies about the "on" and "off" properties. In this way, bare differences are defined circularly in terms of their difference from each other. Moreover, if the *Life* world is pure, we know that there are just no other facts about those properties because we know that the physics tells us everything there is to know. I say the difference is *bare* because it does not rest on any further categorical facts about the properties (if the world is pure). It is a difference that is

ungrounded by any further facts about internal structural differences between those entities or internal relations of difference or contrast between unspecified structureless intrinsic contents.

Postulating facts about intrinsic natures in the *Life* world would violate the purity condition we are working under, because no facts about intrinsic natures are specified by its physics. Thus a *Life* world with any basis other than bare difference would be an impure *Life* world. For now, I think the best way to conceive of a bare difference between two properties, x and y, is to think of the relation as primary rather than implied by other facts, with the existence of the relata, such as they are, derivative on their participation in an ungrounded relation of difference.

Bare differences are difficult to conceive of. Some readers may reject the idea altogether, insisting that a *Life* world must have some kind of intrinsic basis. With an intrinsic basis, there would be a contentful difference where the existence of the difference would be derivative on further facts of intrinsic difference between some unspecified natures of the relata. It is clear, however, that such facts about an intrinsic basis would go beyond what is specified by the bare laws of the *Life* world. I argue later that such an intrinsic basis is crucial to the production of consciousness, but to presuppose it now would beg the question about whether pure physics can specify an adequate basis for the world. So I stay with the "bare" understanding of *Life* for now and examine it more critically later in the book.

My defense of the Skeptic's Claim begins with a closer look at the materials available in the *Life* universe. To reiterate: What does it mean to be an *on* or *off* property? The only two requirements are that (1) they should be distinct and (2) they should be instantiated in patterns conforming to the rules set down by the three dynamical laws. In short, the distinction between being *on* or being *off* is a merely formal one. *On* and *off* specify bare, content-free difference.

Because it specifies only bare difference, the *Life* specification is, at heart, a structural schema for a universe. It specifies certain patterns of contrast between kinds of being, patterns that must hold for a universe to count as a *Life* universe. As we ascend to higher levels of organization in the pure *Life* world, we do not escape from the circle of bare difference. In pure *Life* we have a world potentially consisting of a huge number of simple, bare differences lying side by side, with reliable, regular transitions between them. A *Life* structure is a pattern of bare difference, mere contrast.

Premise 2: Consciousness contains qualitative content. The skeptic claims that we have observational knowledge that consciousness contains qualitative content, involving what are often called *qualia*. The claim that knowledge of qualitative content is observable is critical to the force of the skeptic's arguments. Without it, there is no strong reason to resist performing a *modus tollens* on the conclusion, simply eliminating phenomenal consciousness and its troublesome qualities from our list of explanatory targets. In the following I support the Skeptic's

Claim by providing a direct argument that qualia are indeed observables. By calling qualia *observables,* I mean that they meet four conditions:

1. They belong to a type whose members are potential objects of awareness.
2. We can become aware of them without the aid of special instruments.
3. The dubitability of our belief in facts of the relevant type is almost zero.
4. Our awareness of instances of the type is reliable.

Some people do raise objections to the claim that qualia are observables (e.g., Wilkes 1988; Dennett 1991b; Akins 1993). The most common worry is that modeling our knowledge of qualia on perception is misleading, so people are unsure how we can be observing them. Minimally, opponents sympathetic to these eliminativist worries hold that the knowledge grounding the skeptic's conclusions is highly refined, theoretical, and corrigible.

To these worries, the skeptic replies that the objector seems to have an unreasonably narrow concept of observation. By insisting that something can achieve the status of an observable only if we obtain the information about it through ordinary perception, the objector is making too strong a claim. The objector rules out of court a huge amount of information about consciousness that we have access to and that a theory of consciousness should have to explain. I defend the following argument that qualia are observables:

1. Some thoughts and memories are observables.
2. If thoughts and memories are observable, then the evidence for them is observable.
3. Phenomenal contents (i.e., qualia) provide evidence for observable kinds of thoughts and memories.
4. Therefore, qualia are observable.

As examples of observable thoughts and memories, here are two statements that most would agree express observable facts:

(A) Last night I thought about my childhood.
(B) Sometimes I think about my childhood when no one else is around.

The previously defined characteristics of observables allow facts such as (A) and (B) to attain the status of useful falsifiers for scientific and other theories. For example, a theory of mind fails to account for some of our evidence about ourselves if it fails to account for how we can sometimes think about our childhoods when no one is around.

Facts such as (A) and (B) are no *more* problematic than many other facts we count as observable. Also, they are *introspectively* observable, and the fact that perception does not mediate our awareness of them seems like a red herring. Basically, if *anything* counts as observable, then (A) and (B) must count as observable, too. Our skeptic firmly insists that a science of mind must recognize observables such as these if it wants to be treated as legitimate. Because facts such as (A) and (B) turn out to be no more problematic as observables than are per-

ceptually mediated facts, a straightforward argument delivers the phenomenal qualities as observable also.

Last night, I lacked behavioral evidence that I was thinking about my childhood. I was not writing about it, nor talking about it, nor acting on it. I was, in fact, scouring my bathtub. How do I know what I was thinking about? What was the evidence of my thoughts? I introspectively observed my thoughts, and my evidence was the presence of certain kinds of conscious phenomenal imagery, verbal, imagistic, and kinematic: phenomenal images of childhood scenes, spoken and heard sounds, and remembered emotions. That imagery may have been identical to the thoughts or it may just be a concomitant of thinking that gives evidence for thoughts the way that snow on the ground gives evidence for cold weather.

In either case, my awareness of the conscious phenomenal imagery cannot be considered more doubtful than my awareness of my thoughts. Because the phenomenal imagery is the evidence for such thoughts, it is easy to argue that[5] the sentence (A) has the status of an observation claim only if the phenomenal imagery that is my evidence for it has the status of being observable. Similarly, I obtain my knowledge of *types* of thoughts such as those referred to in sentence (B) from observables only if I also obtain my knowledge of types of phenomenal qualities from observables.

Arguments such as this, the skeptic maintains, establish that we obtain knowledge of what the phenomenal qualities (colors, feelings, sounds, imagery, other sensations) are like through observation. For example, I obtain my knowledge of what the shades of blue look like to me by consciously experiencing them. Consequently, phenomenal qualities are observables (which is not to say observation of them is always either easy or incorrigible). As scientists, we must hold explanations and theories accountable for phenomenal information obtained through observation.

This conclusion does not cross David Lewis's (1995) recommendation that physicalists must deny that we have special, unmediated access to the true nature of qualia. To possess phenomenal information, our skeptic does not need to have a more direct access to qualia than to any other kind of observable. The skeptic is chiefly concerned with the character of the connection between phenomenal qualities, as disclosed through the phenomenal information we *do* have available, and their hidden natures, if they have hidden natures.

Premise 3: Bare difference does not entail qualitative content. Could conscious experience with its qualitative content arise from bare difference? Bare differences within cellular automata are a surprisingly fruitful ground for the emergence of an incredibly large number of kinds of things. *Life* itself can exhibit phenomena of indefinite complexity. For instance, because we already know that *Life* may contain self-replicating phenomena, we cannot rule out that it could exhibit some kind of genuine life. Because *Life* supports the existence of objects that dynamically evolve, it is at least an epistemic possibility that these entities might eventually lead to the existence of animate objects. We also have to hold it

as epistemically possible that these objects might metabolize elements of their environment, act in a goal-directed manner, adapt to be increasingly complex, and generally possess a suite of functional properties sufficient for regarding them as alive.

Given that life might exist, ecologies might exist. Given that ecologies might exist, even economies might arise in a *Life* universe. We can analyze economies into kinds of functional relations between objects within an ecological system, and functional relations are a combination of evolutionary and interactive properties. So, overall, requiring entailments from lower levels to higher levels in a *Life* world does not give us grounds to rule out many kinds of phenomena in it. Nevertheless, the skeptic holds that no pure *Life* world can entail the existence of consciousness or the specific character of its qualities.

The skeptic maintains that facts about bare difference are always consistent with the absence of experience, because qualitative contents are not merely structures of bare difference. If we consider that our taste space, for instance, contains different tastes and that our color space contains different colors, the relevant premise is that these tastes and colors are contents instantiating a structure of difference relations, not structures instantiated merely by difference relations.

Of course the skeptic knows that we can catalogue the differences between different colors and different tastes along relevant dimensions. If we do this, we can surely abstract out a content-free difference structure. The skeptic's objection is to the further move of analyzing conscious qualities into these abstract patterns of difference between them. Rather, our acquaintance with the phenomenal qualities yields information about them as contents occupying slots within these difference structures. Reification of the difference structure as basic ignores the grounding of those differences in each specific case and so ignores the content instantiating those structures. Given this observation that differences between phenomenal qualities are not themselves bare, our question then reduces to whether or not individual qualitative contents such as the shades of green might be constituted by patterns of bare difference.

We can observe that a pattern of differences between colors can produce another color. For instance, a field of tightly packed yellow and red dots may yield an experience of phenomenal orange under the right viewing conditions. However, we can also observe that the shade of orange that results is not produced by the mere pattern of difference. It has to be a pattern of difference between the appropriate colors, thus providing no explanation of color in terms of *mere* patterns of difference.

If we try to abstract the patterns of difference from their contentful bases, viewing colors as mere difference structures, we see that the result is multiply realizable and that some of the realizations do not yield orange. For example, one can instantiate the same structure of differences between two other colors whose hues lie at the same distance from each other as red and yellow (e.g., yellow and green). A pattern of dots of these colors will yield a different color from orange. Therefore, we can observe an identical structure of formal difference but differ-

ent colors. The example shows that, even allowing that we start with colors, one cannot reduce some colors to the mere difference structure among other colors.

The preceding observation is suggestive. After all, the skeptic is maintaining a much weaker position. The position the skeptic is defending is that patterns of *bare* differences do not entail the facts about the phenomenal qualities. Patterns of bare differences are difference structures whose identity obtains because of a mere formal difference, ungrounded by content at all. The skeptic notes that orange cannot even be reduced to the structure of difference between red and yellow once we allow substitutions for the phenomenal content of red and yellow. We can observe more straightforwardly that red and yellow are not constituted by patterns of mere difference, without any content at all.

The skeptic can even recruit Frank Jackson's argument about Mary, the superneuroscientist who spends most of her life trapped in a black-and-white room, to bolster this point. Most find it hard to deny that Mary learns something factual the first time she sees red (even if it is just a fact involving a new mode of presentation for an already known fact). By knowing all the physical facts, Mary certainly had all the information about the patterns of contrast and difference that are relevant to conscious sight. Yet these facts are not enough to yield, even in principle, whatever it is she learns on first seeing red. Whatever one thinks this implies about *physicalism*, it certainly implies something about *phenomenal redness*. It follows inevitably that whatever she learns about the experiencing of red is not just a fact about bare difference or patterns of bare difference. Because those are the only kinds of facts a pure *Life* world could entail, it follows that such a world could not entail the facts about conscious experience.

As an analysis of phenomenal content, the idea that something like a shade of red is a pattern of bare, merely formal differences conflicts with empirical observation. To make it work, something must be added. The only other tool *Life* presents for constructing phenomenal content out of patterns of bare difference is its counterfactual content. Unfortunately, known logics of counterfactuals add only such things as consistency constraints or metrics over similar possible worlds to our tool kit. These are not even the right kinds of things to add to a collection of formally distinct properties to make them add up to properties that are not merely formally distinct. A pattern of bare differences does not become a phenomenal content because another possible world contains a similar pattern or because it is consistent with patterns that occur elsewhere in that same world. Yet that is all we have here. If one tells a skeptic that a pattern of bare differences transitions to another pattern of bare differences, the skeptic can consistently deny that either pattern has to support experiencing. Nothing in the logic of counterfactuals requires that the *transition* should feel like something, either. The *Life* schema thus seems to underdetermine the story about qualitative content. We seem to have good reason for believing that the Skeptic's Claim is true.

Indexicality Indexical facts are facts specifying an honored place in space and time that counts as the center of a world or an honored object that counts as the

reference of terms such as "I" in that world. By being the center of the world, an indexed point or region of space and time provides a point of view from which we can understand the other facts in the world. For example, indexing a point or region as a *Life* world's center would provide the necessary point of view from which we could partition a world's history into past, future, and present; it allows us to partition its spatial coordinates into the place which is *here* and other places that are identified by their distances and directions from *here*; and it allows us to partition the world into physical information that is available at that place (because signals from other places may have traveled to it) and information that is not available (because signals have been lost or have not had time to reach it).

My argument in defense of the Skeptic's Claim has been run without appealing to indexical facts about potential *Life* worlds. Some people believe that facts about consciousness are essentially indexical, and so it would not be possible to derive facts about consciousness from any nonindexical base of facts. This might be true, although I think that is not clear, but it cannot be the fundamental problem with our analysis of the *Life* world. The kinds of information added by including indexical facts are either honorary (this is the *center* of the world), or are relational facts that follow straightforwardly from discursive knowledge of the honorary fact (*this* is true relative to the center), or are elusive in a way that seems quite different from how phenomenal qualities elude entailment by facts of bare difference (*this* moment is *now).* If others wish to insist that adding indexical facts to a pure *Life* world would turn a world of bare difference into a world able to entail facts about phenomenal qualities, then there is some gap between our understandings of indexicality and phenomenal quality that I do not know how to address. Perhaps they have some very substantial theory of indexicality that I cannot imagine and cannot answer, or perhaps their understanding of their phenomenal information is far less substantial than my understanding of mine.

Warning: We cannot transfer information between worlds. When considering the skeptic's argument, we must resist beliefs about our world that might tempt us to smuggle phenomenal facts into the *Life* universe. For example, qualia in our world may perform some functions or correspond to some specific internal structures or processes. I want to emphasize that we must remain constantly aware that the *Life* universe is not our universe. We are to imagine an alien dimension, a dimension fully described by *Life's* physics. No one can decide the question of whether any conscious feeling at all can exist in a pure *Life* universe by an appeal to first-person evidence, analogy, or verbal reports. This takes out of play certain ground-level intuitions that affect the discussion about consciousness in our universe.

For instance, we cannot claim that we will empirically discover that different kinds of descriptions coordinate (Flanagan 1992) in a way that allows us to attribute an identity or determination relation between conscious feeling and the functionality of *Life* objects. Establishing such coordination would require us to access facts of both kinds, and the problem is precisely to access the phenomenal

facts in a pure *Life* world, if any. Conversely, if we had access to the phenomenal facts, if any, we would obviously not need any process of "coordination" between them and other kinds of facts. Those other kinds of facts would have been the entailment base from which we obtained our phenomenal information. So the pure *Life* universe is alien to us, and only entailment could bring consciousness into existence within it.

What phenomenology, if any, would cognitively functioning objects in the *Life* world require us to attribute to them? We do not have first-person knowledge that even one conscious state exists in such a world. Without some supporting story about how the facts in this alien world can be sufficient to support facts about consciousness, we cannot assume the existence of consciousness. And the supporting story must go beyond a coincidence of facts in our world if we want to generalize from our world to a hypothetical *Life* universe. Our alienation from the *Life* world blocks us from transferring the information so naively.

At this point, the existence of an explanatory gap in our world, admitted even by many physicalists, is evidence that mere coextensiveness, or "coordination," is all we really have. If this is so, then the functional information in the *Life* world by itself cannot be the whole story that we would need to attribute consciousness to *Life* objects. It follows that a skeptic is consistent if he admits to any kind of functioning at all in the *Life* world and denies that the activity supports consciousness.

2.6 From Life's Physics to Earth's Physics

If a pure Life world cannot entail facts about phenomenal consciousness, then a pure physical world cannot entail facts about phenomenal consciousness. Cellular automata such as *Life* very closely capture the essential character of our scientific concepts of the physical world and physical properties. In fact, it is not too difficult to imagine that our world might be a giant cellular automaton, albeit perhaps one with complicated stochastic causal-role properties. By using genetic algorithms to discover evolution rules, researchers at the Santa Fe institute have even discovered automata that produce particle-like elements capable of moving from cell to cell and interacting. An automaton can use these particles as information-bearing elements useful in solving problems encoded in its initial state (Das, Mitchell, and Crutchfield 1994). More recently, Wolfram (2002) has reported results of his twenty-year study of cellular automata, arguing from a tremendous amount of data that understanding our world in terms of cellular automata throws light on fundamental and unsolved problems in almost every branch of science, including fundamental physics.

Even if the concept of a cellular automaton does not perfectly capture our notion of the physical world, our concept of the physical is sufficiently close to that of cellular automata that it seems as if the same restrictions apply. They seem to be the same in the relevant respects. In particular, they share a common commitment to bare differences in their fundamental postulates. In *Life* we have *off* and *on* properties. In physics we have *spin*, *color, flavor, charge*, and *mass*.

The theoretical character of the basic properties is just the same in both cases: One stipulates at first that they are distinct and fleshes out their natures by designating laws that describe how they behave. The only real differences between *Life* and physics lie in such attributes as the complexity of the laws, the number and kind of dimensions the cells exist in, and perhaps nonlocal causation. In our world, the structure of the basic entities is more exotic. Instead of squares, we have particle waves and fields, maybe ten-dimensional strings with six of their dimensions rolled into an inscrutable knot, or other such exotica. Perhaps causation in the physical world requires infinite calculation, and so a Turing machine cannot simulate it.

At best, these differences add degrees of vagueness or complexity to the notions of structure, interactive property, and so forth that already are present in cellular automata but that do not seem to make a fundamental saving difference. The failure in the *Life* universe does not seem to arise from the facts that the basic objects were squares rather than strings or that the causal role properties were related simply and locally rather than complexly and nonlocally. Rather, the failure was rooted categorically in the stark geometric and bare counterfactual nature of the properties and of the world they made. Our experienced world is a world of felt tone: warm and warring, whirling and worrying, color and cadence. The pure *Life* world is a ghostly crystal, a home to phantasms.

So it seems that *any* phenomena that the physical facts entail must be analyzable into one of the basic classes of properties or some combination of them. Combinations of these classes support properties such as location, causal role properties, historical properties, structural properties, evolutionary properties, and properties of interaction. Such properties ultimately need nothing more than patterns of bare difference to exist. Again, the complexity of our universe introduces some vagueness into these general concepts, a vagueness that our specific concepts mirror; but vagueness hardly seems like the kind of thing that will allow us to escape the trap.

Some (e.g., Churchland and Churchland 1990) argue that antiphysicalist arguments using thought experiments are arguments from failure of imagination:

> The negative arguments here all exploit the very same theme, viz. our inability to imagine how any possible story about the objective nuts and bolts of neurons could ever explain the inarticulable subjective phenomena at issue.

It should be noted that the argument I have given does not have the form, "I cannot imagine how such-and-such could possibly explain consciousness, so such-and-such cannot explain consciousness." Its form is, "We have reasons for thinking that the physics of *Life* can only entail facts about bare difference and patterns of bare difference. We have observational evidence that the facts of consciousness are not facts of this type. Therefore, we have good reason to believe that a pure *Life* world cannot entail the existence of consciousness." Its essential form is an argument from insight. It argues from facts about a theory to that theory's failure of prediction, making a direct argument from what Chalmers

calls the "absence of analysis" and not an indirect argument from conceivability or new knowledge. To paraphrase Dennett (1991a), a perception of failure is not the same thing as a failure of perception. The successful perception, in this case, is of a failure of prediction. All that is then required to make the move against physicalism is pointing out that a pure physical world shares exactly the same damning characteristic that causes failure in the pure *Life* world.

A possible disanalogy: The intrinsic properties of the physical. One disanalogy suggested by some people is that physical things have an intrinsic nature, whereas the entities and properties in the *Life* world are defined in a way that disregards such intrinsic facts. These facts about the intrinsic nature of the physical, some feel, may be responsible for the existence of phenomenal consciousness. If so, proponents of this objection argue, consciousness would be physical after all. Stoljar (2001) gives an interesting defense of this position.

I am sympathetic to this kind of view and defend something like it in part II. But I believe it fails to salvage physicalism, instead yielding a dual-aspect theory in which nature has both extrinsic physical and intrinsic phenomenal aspects. In short, a logical gap exists between (1) the observation that the proposed intrinsic properties are properties *of the physical* and (2) the conclusion that *they are physical properties*. After all, not all properties of biological things are biological properties, nor are all properties of economic things economic properties. I argue that physics really does specify only the relational network. Because physical theories are committed only to the existence of the facts they specify, the proposal that physical things have an intrinsic character implies that the commitments of our scientific physical ontology incompletely catalogue the world's properties. At best, physical theories may get at the intrinsic base only indirectly by needing it as part of a metaphysical, or at least extraphysical, framework.

One can take the view either that physical properties such as mass and spin are actually relational properties (e.g., if one believes no reference frames are privileged) or that they are intrinsic properties that physics specifies relationally (e.g., if one believes that the rest frame delivers the intrinsic value of mass). In either case, it is difficult to reconcile the view that physics describes intrinsic natures with the hard fact that general relativity is a fundamental physical theory.

Physics describes the outcomes of potential measurements. If measurements of mass were measurements of an intrinsic property, one would expect to find that an instance of it has the same value in all frames of reference. After all, intrinsic properties are the paradigm case of context-invariant properties, and the measurement of something intrinsic will not vary according to the frame of reference from which it is measured. But this is not what physical theory tells us the physical properties are like. According to general relativity, the very same instance of mass (for instance) has different values from different frames of reference. The same is true for some other physical attributes, such as velocity and shape.

If measurements of mass were instead measurements of the potential effective *differences* mass makes to processes in the frame of reference, finding out that

mass measures differently in different frames of reference poses no intellectual puzzles. Nothing is paradoxical about the idea that a dispositional property, mass, may have a different effective impact in different contexts. Even if one believes that "rest mass" provides a preferred intrinsic value for mass, it is most accurate to think of the measurements of mass in physics as the measurement of pure masslike differences that exist relative to frames of reference. This second, relativistic understanding of mass is the bare difference conception: Mass is first something distinct from the other named properties, and it is further differentiated as the property capable of having a certain kind of dynamical impact relative to other properties. It is these things and not any "intrinsic character" to which physics is committed. The methods of physics suggest that the measurement of all physical properties is essentially similar in intent and outcome, even if as a matter of fact some (like spin) happen to be invariant.

This feature of our physical concepts is very deep and fundamental, and it survives revolutions in theory. For example, in his book *Three Roads to Quantum Gravity* (2001), Lee Smolin gives exciting details about the progress being made on the successor theory to quantum mechanics and relativity. Here is how Smolin describes what our new version of the fundamental character of the physical world will probably be like, according to the latest breakthroughs:

> In the earlier chapters I argued that our world cannot be understood as a collection of independent entities living in a fixed, static background of space and time. Instead, it is a network of relationships the properties of every part of which are determined by its relationship to the other parts. In this chapter we have learned that the relations that make up the world are causal relations. This means that the world is not made of stuff, but of processes by which things happen. . . . processes carrying little bits of information between events at which they interact, giving rise to new processes. They are much more like the elementary operations in a computer than the traditional picture of an eternal atom. (p. 63)

This new physics is directly a picture of bare difference (just as the current picture is). This difficulty with the relativity of physical predicates leads directly to another problem with the suggestion that the physical facts are facts about intrinsic properties. It is extraordinarily difficult to show that physical things *must* have intrinsic properties. All that our best physical theories describe is a network of effective dispositions, with each element typed according to its place in a network of relations to other such dispositions. This kind of purely relational world intuitively feels absurd to some, but arguments showing it to be incoherent do not seem at hand. In part II I produce plausibility arguments against the purely relational view. I believe the arguments there give strong reasons for preferring an alternative view, but they fall short of ruling out the relational view altogether. The claim that intrinsic properties must carry the dispositions described by physics ultimately must be added to theory as an intuitively justified metaphysical axiom. As such, it stands out as a primitive further fact relative to our scientific knowledge of the physical.

Finally, Stoljar (2001) distinguishes between the physicalism I have described here—which he calls *t-physicalism* (theory-based physicalism)—and an alternative that he calls *o-physicalism* (object-based physicalism). O-physicalism holds that a property is physical if, and only if, it is a "property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents or else is a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents."

Notice that even if one accepts the existence of some intrinsic character possessed by the basic physical entities, and even if one extends one's notion of physical fact to cover facts about intrinsic character, science is still left with a bootstrapping problem. The bootstrapping problem concerns how to get from (1) the claim that electrons, photons, and quantum chromodynamic quarks have an intrinsic character to (2) the conclusion that there could exist a human consciousness with further intrinsic character all its own to (3) an explanation of why this intrinsic character associated with a middle-level object should be experiential in nature to (4) an explanation of why this experiential context has various features that it seems to have, such as its peculiar unity and coherence with information processing characteristics of the brain. Why shouldn't intrinsic character be delimited right at the boundaries of the microphysical, with the rest of nature, including us, being mere abstraction off the patterns of their interaction? And why shouldn't intrinsic character be *merely* intrinsic, a kind of categorical nature not experienced at all? Even if one agrees that nature needs an intrinsic basis, it looks as if nature ordered an ordinary Volkswagen and God delivered a top-of-the-line Mercedes, which is very odd.

The proposal that o-physicalism really is a kind of physicalism can sound deceptively reasonable at first glance, but the devil is in the details. An o-physicalist could view this book as an attempt to present the details needed to make an o-physicalist view really work. For now, I merely want to point out that we have no reason to believe that we can solve the challenge of placing consciousness without appealing to some kind of new facts about the world, over and above those that science recognizes as the physical facts. Thus I use *physicalism* to mean t-physicalism.

To an extent, whether o-physicalism also deserves the name is a disagreement about labeling. Knowing the end of the story, I believe the o-physicalist path takes us so far beyond what physicalists have traditionally seemed to mean by *physicalism* that it is unreasonable to think that the view that results is physicalist. For readers who consider themselves o-physicalists, I recommend absorbing the full story and then deciding. Even if the specific development of Liberal Naturalism in this book is not ultimately accepted, it is in the same family of theories as those an o-physicalist will have to develop and eventually endorse. Therefore, it helps to make clear the magnitude of the departure from a physics-based physicalism required by such a view.

2.7 Summary

The failure of the *Life* world to entail facts about consciousness seems to be a result of the fact that the phenomenal aspect of experience is not the sort of thing that is entailed by the existence of patterns of bare difference. In particular, it possesses a qualitative content whose existence is not entailed by facts of a functional, structural, or evolutionary sort. Given that, the failure should hold in our world for basically the same reasons it holds in *Life*. Indeed, in many ways the *Life* example gives us a better proving ground for making the determination, because, unlike in our world, we do not start with the knowledge that consciousness exists in that universe. Therefore, the temptation to see entailments where they do not exist is greatly lessened.

3

Physicalist Responses to the Argument against Physicalism

3.1 Introduction

This book is primarily concerned with the positive project of understanding the place of consciousness in the world.[1] But part of the positive project requires taking on the negative task of explaining why orthodox physicalist approaches are unsatisfying. The issues involved are often subtle, and the literature exploring them is large. Inevitably, I have not been able to do full justice to my or other people's views on the matter, but in this chapter I try to present a fair overview of them. The argument against physicalism is an argument against the claim that the physical facts entail the facts about phenomenal consciousness. The form of the argument I presented in the last chapter was a direct argument from the absence of analysis, not an argument from conceivability or new knowledge. Many find it difficult to deny the arguments against entailment, so responses often question the importance of entailment in the first place. These kinds of responses mainly fall into three categories:

1. *Appeals to a posteriori necessity*. Physicalists may hold that the argument in chapter 2 still only establishes the absence of an a priori entailment, and that a priori entailment is not the only appropriate kind of necessity through which the physical facts may determine other kinds of facts.
2. *Appeals to holism*. Physicalists may claim that antiphysicalists couch the debate in terms of a discredited epistemology or theory of meaning.
3. *Warnings about a greater absurdity*. Physicalists may claim that antiphysicalist arguments must be wrong because accepting them leads to terrible absurdities regarding the causal relevance of experience.

In this chapter I examine each of these strategies for responding to the antiphysicalist arguments and outline my reasons for believing that they are inadequate re-

sponses. The upshot is that the antiphysicalist argument in chapter 2 presents a legitimate challenge to how we view nature and provides motivation in later chapters, in which I embrace Liberal Naturalism as an alternative to physicalist naturalism.

3.2 Physical Ontology and Other Ontology

Ontology is the study of what exists, with particular emphasis on the different ways different kinds of things exist. Metaphysicians (and scientists) engage in ontology by constructing or endorsing theories about the world, and we usually say that each theory *presupposes* or *has* an ontology. A theory's ontology sets out the things whose existence we are committed to if we choose to accept the theory as true. In this sense, both false and true theories have ontologies. The difference is that a true theory's ontology is also the ontology of the world.

Physicalism's fundamental ontology is the ontology of physics, whose nature science progressively articulates for us. Physicalism makes a very powerful claim with respect to its ontology. Physicalism asserts a closure condition, saying that a true, complete, and exceptionless theory of the physical tells us all there is to know about the fundamental nature of our world. It is this claim with which the antiphysicalist disagrees.

3.2.1 Entry by entailment

The need to produce nonphysical facts "for free." Presently, the ontology of the physical consists of fundamental fields, particle/waves (or strings and their vibrations), the dynamical properties they possess, and the laws that govern their behavior. Physicalism is a challenging thesis because the world clearly contains things outside of the ontology of physics. Ordinary life acquaints us with tables and chairs, grass and trees, hopes and fears, wind and leaves. We experience colors and sounds, sweet smells and annoying aches. These things have no explicit place in the ontology of physics.

Physicalism's basic challenge is to accommodate the existence of these things that fall outside of its fundamental ontology. If their nature is not explicitly or primarily physical, physicalism must hold that they are at least implicitly or derivatively physical, perhaps in some extended sense of *physical*. Some treatments that explain ways physicalism can extend its fundamental ontology to nonfundamental things are by Jackson (1994), Chalmers (1996), Kirk (1994) and Poland (1994), with Kim's (1993) being perhaps the most comprehensive work on the issue. Here I outline the intuitive idea behind extending physicalist ontology and how it can be captured within a principled metaphysical framework. In doing so I explore, step by step, how physicalism gets backed into its corner.

The following framework is related to that used in Chalmers (1996), with some modifications in substance and presentation. The major departures from Chalmers are (1) in the way that the appropriate mode of supervenience is de-

fined; (2) in the definition of fact used by the framework; and (3) in that with it there is no need to appeal to two-dimensional semantics in any of the analyses of physicalist responses to the antiphysicalist arguments.

The basic intuition behind physicalism is, in David Armstrong's colloquial phrasing, that, given all the microphysical facts about our world, all the other facts are an ontological free lunch. To emphasize this point, I say that physicalism requires *for free* connections to exist between the particular physical facts in our world and the particular facts of higher-level ontologies.

To obtain an ontological free lunch, physicalism must show how the high-level facts of biology (or meteorology, chemistry, etc.) are just physical circumstances under another guise. To express this position, the physicalist needs a theory of appropriate *determination relations*, or, as philosophers call it, a theory of *supervenience*.

Formulating physicalism. Supervenience relations have several variations, and some of these variations imply dualism or pluralism rather than physicalism. The right sense of determination really requires understanding what we mean when we say that it is impossible, in a strong sense of impossible, for the physical facts to be what they are and the other facts to be different. To really express an appropriate supervenience relation, the physicalist needs *modal language*, the language of possibility and necessity.

In my discussion of modality in this section, I self-consciously avoid the terms *metaphysical possibility* and *metaphysical necessity*. As its proponents usually use those terms, a world is metaphysically possible just in case it is conceivable and consistent with empirical facts about essence or identity. Typically, people who use the terms *metaphysical possibility* and *metaphysical necessity* explain them by appealing to certain central examples, such as our having discovered through science that water is essentially H_2O. The claim is that, through this discovery, we came to know that alternative imaginable worlds where water is H_2O are *metaphysically possible* because those worlds are like the actual world in the relevant way. However, we also learn that imaginary worlds where water is something else, like an alien but indistinguishable kind of chemical XYZ, might seem conceivable but are not metaphysically possible because water could not really exist without its essence.

In this way, the metaphysically possible worlds encompass a set of possible worlds that we can learn about based on empirical discovery, even if that set is potentially larger than the set of physically possible worlds. As such, we must establish metaphysical possibilities and necessities inductively through scientific or other empirical investigation, if we can establish them at all. Advocates claim that the existence of *metaphysically necessary* facts show us that the space of possible worlds is smaller than we would have thought without science. If these modalities exist, some things that otherwise seem possible are not *really* possible, some things that do not seem necessary really are necessary, and we discover the details of these things as science progresses.[2] This means that deter-

minations also might be discovered empirically because determination relations involve necessity.

The proper interpretation of metaphysical possibility and necessity is a point of high contention between physicalists and their opponents. Because of the controversy surrounding the metaphysical grade of modality, for now my use of possibility and necessity will not appeal to the metaphysical notions. The right modality for physicalism might come to the same thing as a priori necessity and possibility, or it might come to the same thing as metaphysical necessity and possibility, depending on what the a posteriori portion of the metaphysical variety really amounts to.

In place of the "metaphysical" modalities, I use the terms *ontological necessity*, *ontological possibility*, and *ontological supervenience*. I use this "ontological" modality as a baggage-free placeholder, stipulating up front only that it encompasses the relations that provide physicalists with Armstrong's ontological free lunches. So the A-facts *ontologically necessitate* the B-facts precisely when the B-facts are *for free* relative to the A-facts. When this is the case, we can also say that the B-facts *ontologically supervene* on the A-facts. This means the "ontological" modality might name only one set of relations, or it might be an umbrella term that has a finer analysis into several different kinds of relations. For all we know at this point of the discussion, this might or might not include some kind of metaphysical necessity.

How might a necessary connection fail to be *for free*? Consider *natural supervenience*, a necessary connection that holds when the base facts determine the supervenient facts in virtue of some laws of nature connecting the two. More precisely, a class S of supervenient facts naturally supervenes on a class B of base facts if, and only if, there is no possible world with the same B-facts, the same laws of nature, and different S-facts.

The claim that mental facts *naturally* supervene on physical facts does not secure physicalism. Entities connected by laws of nature do not come "for free" relative to one another. The laws may not be free, either, because on some views their existence is a substantial addition to the facts they connect.[3] Although laws of physics can be included in the base of physical facts, further natural laws connecting physical facts to facts about mental entities would mean that mental facts are strongly novel and fundamentally distinct from the physical facts.

For example, today's physics connects gravity to the other fundamental forces only naturally. Without a grand unified theory in physics, we must consider them as equally fundamental forces. And if we do find a unified theory, the unification will come about because the theory connects them conceptually: Conjoined with appropriate boundary conditions, the deeper principles of the unified theory will entail the existence, differences, and relations between the currently recognized forces. Therefore, natural supervenience is not a kind of ontological supervenience.

To take a different kind of case, by considering abstract entities we can see that Chalmers's logical supervenience relation does not quite imply a *for free* connec-

tion, either. First, it is clear that mathematical truths will be logically supervenient on the physical facts, as there will be no logically possible world that is the same in its physical facts but different in its mathematical facts. It does not follow, however, that mathematical facts are an ontological free lunch relative to physical facts. More pointedly, Chalmers argues that intentional facts are naturalistic facts because they will logically supervene on at least physical plus phenomenal facts. But it does not follow that facts about meaning, any more than facts about mathematics, are naturalistic facts. If meanings are abstract entities, it will be our relations to these entities that determine what we mean or believe, and Chalmers's arguments really just show that the physical and phenomenal facts together are enough to fix these relations. So when considering nonnaturalistic proposals about meaning, such as the interestingly argued and developed view in Katz (1990), it is clear that Chalmers's logical supervenience is too coarse-grained to decide issues about meaning.

These examples show that requiring necessities to be *for free* is not empty, and that a claim about ontological supervenience is a strong and interesting claim. It is a substantial question as to whether a given kind of supervenience is strong enough to be a kind of ontological supervenience. For now I remain neutral about whether the scope of ontological possibility and necessity is determined empirically or a priori or by some combination of the two. With this in mind, we can pattern ourselves after Chalmers by saying:

> Definition 3.1: *Physicalism* is true if, and only if, all the contingent, positive facts in the world are ontologically supervenient on the physical facts, including the physical laws.

Contingent. Facts are facts about contingent entities (entities whose existence is not necessary). Any entity whose existence depends in some way on a historical accident, such as a sperm meeting an egg, is a contingent entity. The *positive* facts are the facts about our world that would remain the same even if our world were embedded within a larger world. For example, the fact that I have two arms is a positive fact because it will be true in any world that contains our world as a proper subset, but the fact that I do not have a guardian angel is a negative fact because it can be reversed in a possible world that is just like ours except that it has guardian angels. These qualifications will not be important to my arguments, but I add them because otherwise it is possible to raise problems for physicalism about things like mathematics, and I want to be fair to physicalism by making it about the natural world and not saddling it with claims about abstract entities.

The master argument. Up until this point I have been using entailment liberally, recognizing that there might be a posteriori entailments. From this point forward, I change my use by restricting what I mean by entailment strictly to a kind of a priori conceptual necessity. Later I treat issues of a posteriori necessity separately from issues of entailment. In what follows, I begin an extended defense of Frank Jackson's (1994) *entry by entailment* thesis, which claims that entailment is the

only way to connect physical facts to other facts *for free*. What I say here is different in detail from what Jackson says, yet I believe it is similar in spirit. I introduce a distinction between narrow and wide facts, and my strategy for defending Jackson's entry-by-entailment thesis is to defend the following argument:

1. If the narrow physical facts do not entail[4] the narrow facts about consciousness, then there are wide facts about consciousness that do not ontologically supervene on the wide physical facts (in this chapter, an argument by cases).
2. The narrow physical facts do not entail the narrow facts about consciousness (from chapter 2).
3. The wide facts about consciousness do not ontologically supervene on the wide physical facts (*modus ponens* 1, 2).
4. Therefore, physicalism is false (3, definition of physicalism).

I defend premise (1) of the master argument in the discussion that begins here and extends through the end of section 3.3 of this chapter. I argue first that an entailment relation between the narrow facts would imply an ontological supervenience between the wide facts, and then I make an argument by cases that the alternative relations do not imply ontological supervenience in the absence of an entailment relation holding between the narrow facts.

Two kinds of facts. Consciousness presents an interesting problem because we possess information about it, especially about phenomenal properties, and we possess information through the physical sciences about the physical world, too. Because we have these different kinds of information, we are presented with a task, which is to understand how to place both kinds of information within the world and relative to one another. It is natural to think of this task as the task of trying to properly account for the *facts* we possess.

There are different ways of carving up facts about the world. We can carve them up in terms of their cognitive or epistemic relations, understanding that these may often reflect aspects of things (i.e., a partial way something presents itself to a point of view), or we can carve them up in terms of the external objects and properties they refer to. For example: Do the claims "Clark Kent is a reporter" and "Superman is a reporter" express the same fact? It is natural to say: in one sense yes, in another sense no. I capture this difference by saying that the two sentences express different *narrow facts* but the same *wide fact.*

In general, we can say that two sentences express different *narrow facts* when they are cognitively distinct. In particular, sentences A and B express different narrow facts when a priori reasoning alone cannot tell a subject that if A is true, B is true, and vice versa. On the narrow conception, facts are epistemically fine-grained, because they reflect the way subjects represent objects and properties in the world, including cases in which just their aspects are represented. That is, narrow facts are sensitive to *modes of presentation* of objects and properties in the world. A consequence is that narrow facts are *hyperintensional:* Two narrow

facts may be distinct even if sentences that express them are true at all the same possible worlds.[5]

For example, the narrow facts expressed by *Superman can fly* and *Clark Kent can fly* are different facts by this criterion. We can prove this to ourselves by imagining their cognitive significance for Lois Lane. Lois Lane can have a perfectly competent understanding of *Clark Kent can fly* while believing it to be false, even if she believes that Superman can fly and is a perfectly competent reasoner. This implies that whatever it is she understands when she understands *Superman can fly* does not imply in the relevant sense that Clark Kent can fly. Similarly, we can see the difference in what Lois understands by observing the huge difference in her view of the world and the accompanying behavior that results if she learns that both facts hold as opposed to just *Superman can fly*. So, if those sentences expressed facts, those facts would count as distinct under my usage of "narrow fact."

Similarly, *Water is liquid at room temperature* and H_2O *is liquid at room temperature* express distinct narrow facts. We can prove this to ourselves by noting that a competent reasoner can have an understanding of the first sentence that cannot be faulted by others even if they know nothing about the chemical composition of water. Additionally, that understanding would not be undermined if it turned out that chemical theory is radically false. Thus identical objects or properties may yield distinct facts if those objects or properties present themselves to a subject in different ways due, for example, to the subject having access to different aspects.

By contrast, wide facts are individuated in terms of objects and properties in the world. We can say that two sentences express different wide facts when they concern different objects in the world or attribute different properties to them. On the wide conception, facts are epistemically coarse-grained. Two statements or beliefs can express the same wide fact even if they are cognitively quite distinct from each other.

For example, the wide facts expressed by *Superman can fly* and *Clark Kent can fly* are the same by this criterion even though the corresponding narrow facts are different. Both attribute the same property (flying) to the same individual in the world. Likewise, *Water is liquid* and H_2O *is liquid* express the same wide facts. These are both cases in which two sentences express the same wide facts but different narrow facts.

My strategy here is to allow both kinds of facts and to concede to the physicalist that there must be a gap in wide facts for physicalism to be false, while nevertheless couching much of my discussion in terms of narrow facts. Casting things in terms of narrow facts allows me to capture epistemic distinctions and to capture the idea that physical facts and facts about consciousness at least involve different aspects or different modes of presentation. In doing so I do not beg any questions against the physicalist, however. I explicitly argue that the absence of an entailment between narrow physical facts and facts about consciousness *implies* that the set of wide physical facts is incomplete.

I suggest that there is a deep connection between narrow facts and wide facts. I argue that, in general, any incompleteness in a set of narrow facts implies incompleteness in our knowledge of the corresponding set of wide facts. In particular, I argue that if one has a complete set of narrow facts, the wide facts will be constructible out of them, and vice versa. The underlying principle is that physicalistic individuals and properties should hide no essence distinct from the sum of all their aspects.

This last point is the most critical. Let F_P be the base of wide physical facts in the world and f_P the base of narrow physical facts, according to the completed ideal physics. By the physicalist hypothesis, knowing the narrow facts in f_P means having complete empirical information about the wide facts in F_P. I defend the thesis that under these conditions a gap between narrow facts implies that we are missing some wide facts.

At worst, the facts in f_p differ from the facts in F_P by entailing the truth about them in a piecemeal way through the various aspects they manifest in different circumstances. This means, for example, that one property might be represented in two ways, just as it is possible that one underlying property might be represented by both mass and energy in today's physics (i.e., there is one property in F_P and two apparent properties in f_P).

The key is the scope of the claim about f_P. Because f_P is claimed to be complete, the wide facts in F_P should contain no more empirical information[6] than is conveyed explicitly or implicitly in f_P. Motto: The wide facts in F_P can hide no essence distinct from the sum of all their aspects, and their aspects are revealed through f_P. For example, this means that laws relating energy and mass to other things and to each other reveal all there is to know about the property of which they might be aspects. The reverse motto is also true: The narrow facts in f_P can convey no information about an essence not found in F_P.

If it holds, this equivalence of information in F_P and f_P creates a bridge between ontological questions about F_P and epistemic questions about f_P. In particular, if a set of narrow facts expressed in the language of a complete physical theory fails to appropriately determine some narrow facts about a thing, then it is leaving out some information about some of the aspects of that thing. Aspects themselves are ontologically expensive: A difference in aspects always involves a difference in the objective properties of a thing.

When a thing is presented to a subject under multiple aspects, there needs to be at least (1) an objective explanation of the connection between those aspects; (2) an objective explanation of a point of view to which these aspects are presented; and (3) an explanation of why the thing shows this aspect from the identified point of view.

To give such an explanation, we need to give an objective story about properties of objects and observers in the world, showing how the object can present the two aspects to an observer. This story itself involves wide facts about the world. The wide facts required to meet requirement (1) must involve wide facts not already accounted for in the original set of facts. The wide facts required to

meet requirement (2) will involve at least that, plus some wide facts about an entity that provides the essential point of view. The wide facts required to meet requirement (3) will express some properties of the thing that it does not exhibit from other points of view and, therefore, some new wide facts about the thing.

Consider energy and mass again. There is a connecting wide fact about the role velocity plays in creating the apparent differences between them (a type 1 wide fact). There are wide facts about the relativity of velocity to inertial frames of reference (a type 2 wide fact). And there are wide facts about the sameness of effects energy and mass have on things in these inertial frames of reference, when understood appropriately (a type 3 wide fact).

Also consider the distinction between the evening star and the morning star. Each distinct concept represents a different aspect of the planet Venus, and one cannot explicate the differences without appealing to new wide facts. Particularly, we need:

1. Type 1 wide facts: facts about the orbit of earth through the solar system, relative to other solar bodies. This helps connect the position of the morning star in the sky to the position of the evening star in the sky.
2. Type 2 wide facts: facts about observers on earth who occupy certain areas on its surface and are sensitive to light.
3. Type 3 wide facts: facts about the rotation of the earth and its effect on how objects in the sky look to these observers.

In the end, the distinction between fine-grained narrow facts and coarse-grained wide facts helps to frame the discussion in a way that is clear and respectful of the differences between interested parties but that matter little to the conclusion of the analysis. The epistemic claim about the fine-grained facts is scoped so ambitiously, saying they reveal *complete* empirical information about the wide facts, that the epistemic facts and relations can essentially stand in for their ontological counterparts.

Because of questions some raise about the epistemology of phenomenal concepts (e.g., McGinn 1989; Loar 1990; Hill 1997), it is important to always keep in mind the scope of the physicalist claim when asking and answering questions about how critical are the entailment relations between narrow facts. For example, Christopher Hill argues that the special character of phenomenal concepts will make every physical constraint we know about, or could know about, *seem* consistent with the possibility that the physical processing could occur without consciousness, even though that is not really the case. This position maintains that in reality there is one property P that is the property of being conscious and that it is understood through two specially independent concepts, one theoretical and the other introspective. Hill's kind of view is an especially popular response to Frank Jackson's Knowledge Argument and is sometimes used to explain away conceivability intuitions. (*Note:* the argument given in the last chapter was not a form of the conceivability or knowledge argument.)

If true, Hill's view might explain the seeming lack of entailment. However,

even if this view were correct, the necessity connecting the physical facts to the phenomenal facts would still need a basis,[7] raising the question of what is being overlooked. Before accepting a Hill-Loar-McGinn kind of view as an answer to the antiphysicalist arguments, we should ask questions about the basis of the a posteriori necessity that is supposed to connect the two kinds of (narrow) facts despite the seeming lack of entailment. For example, analysis may reveal that a position like Hill's wrongly assumes that an empirical identity can be the basis and do the work physicalism requires when it really cannot. Given the scope of the physicalist claim about the narrow facts, I can see only three options for a physicalist responding to the failure of entailment to connect the two types of narrow facts:

1. *A posteriori necessity.* Physical theories are complete specifications of the natures of things, and the necessity is an a posteriori metaphysical necessity.
2. *Opaque entailment*. Physical theories are complete specifications of the nature of things, and there is an a priori entailment from physical facts to phenomenal facts, but we, the theory makers, must inevitably suffer delusions about what they can entail.
3. *Incomplete physics.* Physical theories will always be incomplete specifications of the physical nature of things.

A posteriori necessity. Strategy 1 maintains the completeness of physical theory (or, really, that it can be completed in the ideal limit) and appeals to an a posteriori necessity. This strategy works only if there actually is some form of a posteriori necessity that does work equivalent to the work that entailment does and is also not vulnerable to the antiphysicalist arguments. In section 3.3 I argue that a posteriori essentialism, empirical identity, and the necessity of natures cannot do this work. That leaves the relevant basis for a posteriori necessity still unique and unexplained. So, if followed, this strategy would assume that an appropriate form of a posteriori necessity exists and in no way accounts for it. This is not solid ground for physicalism. I ultimately end my discussion of a posteriori necessity by making a case that, unless we can successfully identify a basis, this strategy is based on an untenable, because empty, claim.

Opaque entailment. Strategy 2 claims that an entailment between the two kinds of narrow facts actually exists but that, even if we had a completed physical theory, we would be congenitally unable to extract it. It could be that the two families of concepts involved are from such alien faculties that there could not be appropriate connecting concepts. For example, a proponent of a Hill-Loar type of view could hold that the phenomenal information truly is contained in the physical facts but that the in-principle independence of physical concepts from our introspective concepts prevents us from ever being able to see how. Or perhaps there could be connecting concepts, but we just do not possess them and never could (McGinn's view).

A strategy of opaque entailment would be viable if the antiphysicalist arguments were arguments from ignorance having the form, *I cannot see how the physical facts could entail the facts about experience, so they do not*. But this is not the form of the argument. As I explained in 2.5, the form of the argument is actually,

> We have reasons for thinking that the physical facts can only entail the type of facts that are constituted by patterns of bare difference. We have observational evidence that the facts of consciousness are not facts of this type, so we have good reason to believe that the physical facts cannot entail the existence of consciousness.

On its face, this argument expresses an insight, not a cognitive blind spot. The scope of the negation is within the knowledge claim ("We know that the p-facts do not entail the c-facts."), whereas opaque entailment best explains a claim in which the negation is outside the knowledge claim ("We do not know that the p-facts entail the c-facts"). To attack an argument from insight by appealing to an opaque entailment, one needs to propose more than a shortcoming in our cognitive ability due to different kinds of concepts. Beyond that there must be some functional glitch in our normal capacities, a glitch responsible for creating delusions of reason or observation, and causing us to perceive a shortfall of empirical information in a set of facts when that shortfall does not really exist. Otherwise, the appropriate response to a position like Hill's, who proposes that the physical and phenomenal facts involve concepts belonging to two highly independent faculties of knowledge, should be to assume that the different faculties are attuned to fundamentally different aspects or properties of things.

To succeed, a position such as opaque entailment will have to finger some error in the *Life* argument, even if we will not be able to see positively just how it is an error. Because *Life* is a simple and clear creation of our own, it is not likely that the problem is our identifying the fundamental facts of *Life* as consisting in bare differences. A more likely route for a proponent of opaque entailment would be to explain our delusions by holding that phenomenal qualities really are structures of bare differences. Perhaps phenomenal qualities are bare differences understood in an "indexical" manner from a "point of view."

Such a diagnosis would attack the core elements of the concept of consciousness and the viability of the idea of phenomenologically valid information. In the end, I fear that it would look much more like eliminativism than nonreductive physicalism and that it would force McGinn's, Loar's, and Hill's ideas to be much closer to Daniel Dennett's ideas than they probably would like them to be. So opaque entailment would only plausibly explain why we might fail to see an existing entailment, falling well short of plausibly explaining why we succeed in seeing a failure of entailment even though the entailment is there.

Incomplete physics Strategy 3 bites the bullet and blames the potential theories, concluding that they are irredeemably incomplete. This response creates problems because physicalists typically answer the question, "What is physical?" by appealing to the authority of physics. Because the physicalist draws the bound-

aries of the physical by appealing to theory, incomplete physics will not work as a defense. The *physical* facts just are those facts to which our ideal physical theories end up committing us. Either the information missing from them is about some extraphysical aspect of the brain or world or the physicalist must produce a theory-independent and non-question-begging definition of the physical. Taking the first route implies that physicalism is false, and taking the second route seems to imply, for the sake of consciousness, that physical science cannot be completed. To me, this second route seems hardly more like traditional physicalism than does the first route. More important, this second route threatens to make physicalism trivial, perhaps little more than the rejection of substance dualism.

3.2.2 The possibility of entailment

According to the entry by entailment thesis, entailment occupies a strategic place in the physicalist claims. The chief challenge for physicalism is to show how a fundamental ontology consisting of the basic properties of physics, such as mass, charge, spin, flavor, and color, possessed by basic individuals, can produce properties and individuals in wholly different ontologies *for free*. By "for free," I mean without introducing anything else fundamental. In section 3.2.3 I explain how entailment answers this need: It provides a necessity through which we can see how one set of facts can determine another set of facts using nothing further except interpretive and conceptual resources. Therefore, entailment is a priori in a way that does not require introducing any new fundamental individuals, properties, or laws into the determination procedure.

In this section, I answer the complaint that entailment cannot be essential to physicalism because there are many higher-level things that are clearly physicalistic even though the physical facts *entail* very little about them. From the antiphysicalist side, Chalmers (1996) and Jackson (1994) address this complaint by sketching examples of how one might establish entailment in particular cases.[8] They also offer some general reasons to believe that entailment holds almost universally. From the physicalist side, Kirk (1994) also defends the claim,[9] as do Horgan (1984) and Armstrong (1982). For more detail regarding the positive case, I refer interested readers to these authors.

I do not repeat the positive case here, but instead directly address the basic viewpoint motivating the objection. People's doubts come from the family of concepts surrounding entailment, such as *logical*, *consistent*, and *analytic*. People view entailment as a logical relation, and they think the requirement that higher-level facts follow *logically* from lower-level facts is too strong. Some believe that entailments need to be "analytically true," where analytic truth requires the ability to produce a syntactically well-formed deduction from definitions. Were this true, the basic physical facts could *entail* supervening facts only if the concepts expressing the supervening facts could be defined using terms from physics, because definitions would be needed to analytically derive the supervening facts from the basic physical facts.

For example, Ned Block and Robert Stalnaker (1998) seem to assume the definability requirement in their extensive criticism of Jackson and Chalmers. In more than one place they say such things as:

> Perhaps the semantics of "water" is more like [a regular proper name] than it is like [the proper name of a definite description], in which case there is no way to fill in the details of "the water role" so that it is a conceptual truth that water occupies the water role. And of course it is even more doubtful that any such analysis of the water role would be both a conceptual truth and be an analysis in *microphysical* terms.

Block and Stalnaker are asking for an analysis of "water" into microphysical terms, which is basically a request for a definition in those terms. Most high-level facts will fail to meet their condition for one (or both) of two reasons: (1) most of the high-level concepts at issue will evade perspicuous definition altogether; or (2) the definition will not be expandable into purely physical terms and so will not be in a form appropriate for use in a derivation from a purely physical base of facts.

I will argue that these worries are beside the point. Entailment, analyticity, and consistency are part of the semantics of thought. The logical systems that inform our modern understanding of them are merely *theories*. These theories have been useful in limited ways but are far from exceptionless. Also, they rest essentially on the ideas that sentences adequately represent thoughts and that the relations between thoughts correspond to the syntactic transformation of sentences according to rules. We currently have no reason to place more than slight faith in these ideas, as the semantics of thought seems to be far from perfectly represented by the theories.

By contrast, I think we have strong reasons for believing in entailments of the relevant type, and these reasons are prior to the acceptance of any logical framework for the analysis of thought. If this is right, then the grounds for skepticism about the relevant kind of entailments are very weak. Properly interpreted, those reasons provide a stronger warrant for being skeptical of the theories than for being skeptical of the relevant kind of entailments.

The tools we use to build our theories, just like the theories themselves, must answer to the informal competence we possess (or may develop) with the use of our concepts, with meaning, with consistency, and with logical consequence. Only if they pass these tests may we accept them as completely adequate theoretical tools. Thus to raise these skeptical objections against entailment requires defending some very strong claims about the adequacy of syntactic derivations in capturing informal semantics. These claims are that (1) the syntax of logical definition adequately models the structure and behavior of meanings for nonprimitive concepts and (2) the logical satisfaction of concepts is adequately analyzed within formal logic. I argue primarily against claim (1) because I believe that claim (1) provides crucial support to claim (2).

If claim (1) were true, concepts without definitions would not be able to support implications and entailments at all, yet they can. As an example, consider the

concept of "friendship." "Friendship" expresses a meaningful concept whose meaning is not primitive, and "friendship" is plausibly vague enough to be unanalyzable in terms of a formal definition involving other concepts related to one another by the logical constants. If this is true, as it seems to be, then it is trouble for someone needing to defend claim (1). It seems to suggest that meaning and implication outrun the resources of formal definition.

Linguistic intuition strongly suggests that notions of entailment and satisfaction apply to "friendship." When one considers concrete cases, some situations certainly present themselves as being entitled to the claim that they *conceptually* satisfy the meaning of friendship, irrespective of its resistance to formal definition. To see this point, imagine two people, Allen and Gregg, who have known each other for fifteen years, genuinely like each other, have shared many experiences, secrets, and adventures, go out of their way to be in each other's company, and rely on each other for advice and support in times of stress. What would we say to someone who, in full knowledge of these facts, nevertheless claimed that Allen and Gregg were not *friends*? The most straightforward and sensible answer is to assert that this skeptic does not understand what it *means* to be a friend.

In any case, the dialectic would hinge on an adequate grasp of the concept. If the skeptic produces contravening evidence, it will be contravening precisely because we recognize it as inconsistent with the *meaning* of friendship. Without that evidence, we will conclude that the skeptic simply does not understand the concept and therefore cannot see what is apparent to everyone else: that Allen and Gregg's history together, and feelings toward each other, entail the fact that they are friends. Our capacity (or incapacity) to represent the meaning of the concept in a formal logical system plays no role in establishing the ability of certain conditions to imply facts involving it. This example is contrary to something Block and Stalnaker claim. Using the example of life, they write,

> More relevantly, it is doubtful that fulfilling any set of functions is conceptually sufficient for life. A moving van locomotes, processes fuel and oxygen and excretes waste gasses. If one adds a miniaturized moving van factory in the rear, it reproduces. Add a TV camera, a computer, and a sophisticated self-guiding computer program, and the whole system could be made to have more sophistication, on many measures, than lots of living creatures.

What does Block and Stalnaker's example show? It is meant to suggest that there are no sufficient conditions that entail facts about "life," but it surely does not show that much. What if the van's materials were synthesized from organic materials? What if it executed competitive strategies for obtaining its fuel at the expense of other vehicles? What if it had a life cycle of self-organized growth, differentiation, self-repair, and deterioration? What if the van factory in the back sprang up as part of this life cycle? At some point the question of its being a synthetic life form would surely raise its head. That could not be a *purely* empirical question because we would already have all the relevant empirical information. After considering what we know about the van, we would have to consider just

what we mean when we say that something is alive and decide whether or not the concept applies. That is, we would consider the broadly logical conditions for satisfaction of the concept.

This decision process would use empirical information, but it would also involve the assimilation and organization of this information into a previously existing and useful category, the category of *living things*. It would be a process, part social and part rational, of either discovering or establishing an entailment. Whether discovery or establishment would occur depends on what action is required to settle the case. If settling the case would require sharpening a previously vague boundary or moving a boundary, then we would be establishing a new entailment. If it instead involved explicitly discovering and recognizing previously unrecognized but sharp conditions of satisfaction, then it would be a discovery.[10]

In fact it is easy to construct examples that show how a physical situation can entail the existence of something that is not defined, nor definable, in microphysical terms. Consider a simple device with two states, state A and state B, and two inputs, input *a* and input *b*. It computes as follows: Whatever state it is in, if it is input *a*, it goes to state A. If it is input *b*, it transitions to state B. We can see this machine as a simple recognition device: It reliably recognizes its two inputs, and the abstract machine is neither defined nor definable in physical terms.

Now imagine a trick lock that is always either locked or unlocked and two keys for it, a gold key and a silver key. Regardless of what state it begins in, if you put the gold key in, it will lock, and if you put the silver key in, it will open. The locking system clearly implements the simple recognition device, and we need only a description of the two systems to see this. Consequently, the implementation relation here must be a priori. Indeed, it is an entailment from the structural facts about the physical situation to the structure of the formal machine, where the former are clearly logically sufficient to ensure the existence of the latter. It follows that definition in microphysical terms has little to do with a priori entailment.

So "friendship" provides an example of an entailment without definition, and the recognition device is an example of entailment with definition, but not definition in microphysical terms. Examples like these can be easily found, and so it does not seem that the proposed understanding of entailment in terms of analytic definition is workable.

Another passage in Block and Stalnaker also provides an illustration of how science relies on informal competence to recognize entailments. They suggest a way that the concept of life came to have its extension:

> There are some paradigm cases of living things, including some that are quite simple. (We need not assume that even the paradigm cases of living things are alive). We understand completely how some of the simpler forms of life work. We have reason to think that more complicated living things work by similar principles, and see no bar, in principle to extending our explanations of simple living things to all forms of life—closing the explanatory gap in the case of life has nothing to do with

> any analytic definition of "life", but rather is a matter of showing how living things around here work.

As a general repudiation of the relevance of conceptual truth to the explanation of life, this account faces a serious problem. If Block and Stalnaker are right, how do we know that those simple creatures we "understand completely" are alive? According to their account, these things are all true:

- None of the facts we know about these simple things show conclusively that they are alive, due to not having any a priori connection to our concept of life. The conclusion that they are alive is a kind of hypothesis.
- The fact that they are paradigm cases of living things does not guarantee that they are alive. Paradigm cases of living things may not be alive.
- A complete understanding of how they work explains why they are alive.

Notice the apparent tension between their third point and their earlier claim that no amount of functional information is sufficient to entail that something is alive. Worse, if the first and second points are true, then the third point really is uncertain. Even after we have all the physical facts about how simple creatures work, we are just making a posttheoretical conjecture that they are alive. It follows that everything we know is consistent with those creatures *not even being alive*. If they are not alive, explaining how they work certainly does *not* explain life. Because conceptual connections are inconclusive, this raises the "scientific" question, *What basis do we have to believe those simple creatures are alive*? If this question is not answerable, and answerable by appeal to "how they work," then we do not really have an explanation of life. Finally, what kind of basis do those things provide for deciding whether these simple things are alive if not a conceptual basis?

If Block and Stalnaker were correct, we would have to leave it open that we have no explanation of life, not because our explanations are empirically inadequate but because what they explain may fail to establish the presence of life. Worse, we could apply the same sort of argument to ourselves. If their first and second points hold, we cannot even determine conclusively that *we* are alive! After all, no mere accounting of *other* sorts of facts about us can settle the question of whether we are alive, as that would require an entailment from those facts to our concept of life. The fact that we are a paradigm case matters little either (for some people, dolphins may be paradigm cases of fish, but that does not make dolphins into fish). Just as with a worm, a completed biology will have to make an inductive guess that human beings are, in fact, *alive*. If other facts cannot entail the facts about life, life itself eludes our grasp. It becomes a mere will-o'-the-wisp that science chases by making inductive guesses about its presence after (or worse, *before*) learning other facts about things. A position that makes room for inductive uncertainty about whether we are alive, even after we learn all the other facts about ourselves, does not do justice either to linguistic intuitions or facts of practice.

Along with Jackson and Chalmers, I believe Block and Stalnaker are overlooking some practical constraints on reference determination. For referential concepts to have value at all, people must be able to determine the reference of their concepts from other facts. Otherwise, we would be in a constant state of futile uncertainty about how our thoughts connect us to our environment.

This oversight is shown when philosophers, including Block and Stalnaker, try to use *thought experiments* about reference in different possible situations to make their points. These thought experiments reveal conceptual truths. At their best, they describe a situation in a way that allows us to determine what the reference of various terms would be in those situations *by recognizing, from a description, what the concept would refer to in the situation*.

Consider the classic thought experiment that is supposed to show the a posteriori character of reference, Putnam's Twin Earth. Briefly, we are to imagine a world macroscopically indistinguishable from earth, but with a different liquid, XYZ, playing the role of water on that planet. We are also to imagine that I (or some person from our world) have a "twin" on that planet, defined as someone who is physically identical to me in relevant respects. We are supposed to discern that, when he uses the word "water," he refers to XYZ; when I, perhaps his molecule-for-molecule duplicate, use the word "water," I refer to H_2O. It is supposed to follow that our "water" concepts must share the same a priori features, because we are internally just alike, yet the term's references are different, showing that a priori features cannot determine their references. Instead, reference is dependent on external factors.

How, though, do we know that the concept of my twin on Twin Earth really does refer to XYZ rather than H_2O? If given the analogous thought experiment, wouldn't my twin know that my concept refers to H_2O? This seems to show some aspect of meaning our concepts share in common: the ability to take us from an epistemic scenario where our concepts are applied to referents for them in the scenario. Chalmers calls this portion of meaning a concept's *primary intension*.

While Block and Stalnaker claim we have no reason to believe in such an intension, their own favored examples show that they must be wrong. All we have in their examples are descriptions of physical situations. Were Block and Stalnaker right, no such description could suffice to tell us anything about reference and nothing in the situation would contradict the idea that my twin's concept refers to the same thing that mine does. Perhaps his concept, like mine, would refer to H_2O after all? Or perhaps my concept refers to XYZ, and I have just been led astray about its content by the highly plausible but nevertheless false belief that water is a substance in *my* world? There would just be no way to really know.

Obviously, this uncertainty does not really exist. We can determine the reference of our concepts a priori, given enough information about the situation in which they are being used. When Block and Stalnaker give counterexamples to proposed definitions of "water" or "life," they are actually demonstrating the ex-

istence of the a priori function that they are arguing against. The evaluation of every example requires using the machinery they think we do not have or else the examples are worthless.

Together, all the preceding points make a strong case against the adequacy of formal or other linguistic definitions as tools for fully representing meaning. If our current formalisms fail in that, then it follows virtually immediately that formal logic has not adequately represented *satisfaction,* either. From that, it follows that we also do not have an adequate *formal* representation of entailment or analyticity. These conclusions should not be big surprises. They are just what one should expect, given that we have only the vaguest ideas at the moment about what exactly concepts are, or how they are structured, or what meaning itself is. These issues lurk in the metatheory of all sciences and in all argumentation, and they do not present special problems for the arguments used in the science and philosophy of consciousness.

The antiphysicalist arguments depend on applied competence at the object level, not misapplied metatheory. Any theory, argument, and explanation will use concepts of one sort or another and will rely on our ability to understand relations between them. Unless it turns out that the antiphysicalist arguments are relying on concepts and meaning in a way that is different from scientific explanation and understanding generally, then these open questions within the philosophy of language, although immensely interesting and important, will not be especially germane to the issues surrounding the truth of physicalism.

3.2.3 Defining entailment

If physicalism is true, then there must be a way to connect narrow facts about the microphysical world to narrow facts about the macrophysical world, and this method of connection must be consistent with the restrictions of ontological supervenience. To meet these restrictions, the bridging principle(s) that enable the physical facts to determine other facts must work in a way that is *for free*. I have been maintaining that *entailment* between narrow facts provides a kind of *for free* connection. My strategy is to substantiate in this subsection the claim that an entailment relation between narrow facts implies a *for free* bridge between wide facts and then, in section 3.3, to explore the alternatives, showing why each alternative fails.

For the physicalist, the problem cases involve facts that are not directly facts of physics. By definition of *narrow fact,* these must be facts that involve one or more concepts that are not concepts within basic physics. As I argued above, it is reasonable to suppose that these problem concepts have necessarily sufficient conditions for their application, therefore it is reasonable to suppose that sometimes facts[11] may hold because more fundamental facts satisfy these necessarily sufficient application conditions. Often these application conditions will be structural, functional, contextual or historical in character, or some combination of those. So even if a narrow fact involves a concept not in the domain of physics,

it may be possible for the wide physical facts to determine these facts *for free* if the narrow physical facts satisfy the application conditions of the concepts in the new domain.

I regard the act of determining whether a situation satisfies the application conditions of a nonprimitive concept to be a kind of interpretation. For an interpretation to be consistent with physicalism, it must meet very strict standards and cannot merely rely on circumstantial evidence. It must be a kind of function that takes the physical facts as input,[12] which maps them onto one or more of their aspects, and decides whether the mapped-to aspects satisfy a necessarily sufficient condition on the relevant concepts in the nonfundamental domain. In performing its operations, the interpretation function cannot introduce any new empirical facts that are not themselves derived from the physical base. Acceptable transformations will be structural, contextual, or logical in character: They will merely preserve physical information, reduce or compress physical information, or show logical consequences of that information.

A little more formally, a base of physical facts *B* may satisfy the application conditions for concepts from some supervenient domain *S* either (1) directly by application of an interpretation function on B that satisfies applications conditions belonging to the concepts characterizing S or (2) indirectly, by an application of an interpretation function satisfying the application conditions of concepts in another supervenient domain *S**, where the *S*-concepts have application conditions that can be satisfied by an interpretation function applied to the *S**-facts. This relation is a kind of *realization via interpretation*.

For example, consider the case described previously of the simple recognition device implemented by the lock that opened and closed depending on whether a gold or silver key was inserted. A necessarily sufficient condition for being such a recognition device is to support a one-to-one mapping to the tokens, states and behavioral relations included in the machine description. Given the description of the physical situation, there is clearly an interpretation function that takes us from that physical situation to a guise that supports such a mapping. Therefore the physical situation realizes the abstract machine via interpretation. In general, this shows one way a physicalist might be able to truthfully assert statements such as *A table exists,* even though the category *table* is not part of the fundamental physicalist ontology. The physicalist could accommodate this ontological novelty by saying that the physical conditions in our world, at some region of space-time, realize via interpretation the property *tableness* by having a guise or aspect that satisfy necessarily sufficient application conditions on the concept *table*.

Note that *interpretation* as defined is a priori.[13] Therefore, the truth of any actually realized facts would follow a priori from their base facts whenever they are realized via interpretation. Before moving on, it is worth noting that the idea of a situation satisfying a concept's application conditions does not assume that application conditions are classical necessary and sufficient conditions. It is perfectly consistent with this account if conceptual structure and conceptual behavior is more complex than definitions of necessary and sufficient conditions can capture.

In particular, the explanation of how a situation satisfies a concept's application conditions may invoke fuzzy similarity rules, distance from prototypes, or cross-modal conditions invoking image matching and application of motor schemas or may produce categories of family resemblances.

In the rest of this book, this sort of connection between sets of facts is what I mean by an *entailment*. An entailment between two sets of facts exists whenever the bridging principles that take us *from* the antecedent facts *to* the consequent facts are entirely interpretive, involving nothing but the correct application and needed refinement of appropriate concepts to a given situation (e.g., a physical situation). In general, I am going to treat an entailment relation as a kind of *informational containment relation*: The antecedent facts entail the consequent facts because they contain all the *empirical* information input into the act of interpretation. This treatment is sensible given that physicalistic interpretation is a function that accepts only physical facts as input and maps them to outputs without adding any fundamentally new empirical information to its input.

An important note: This containment condition is different from a Kantian kind of claim that analytical connections obtain when one concept contains another. For example, if distinct concepts connect to one another by relations of activation and inhibition, the very idea of one concept "containing" the other would not make sense, yet the preceding explanation of entailment as containment would not be affected. In any case, under this treatment of entailment, the fact that something is red entails that it is colored, but not necessarily because the concept of red contains the concept of color. Rather, it is because the fact that something is red contains enough empirical information to determine that it is colored *if one has the concept of color*. The notion of entailment itself is neutral on the relation between the concepts of red and color and does not require containment between the concepts.

The ontological innocence of entailment should be clear. If the base facts, say the physical facts, *contain* all the empirical information needed to establish the truth of the consequent facts via interpretation, then those consequent facts clearly cannot be anything extra. *Containment relations are parsimonious.* It is clear that entailment so defined is a kind of ontological supervenience:

> *Necessity*: If the *B*-facts contain the empirical information constituting the *S*-facts, then any possible world containing the *B*-facts will contain the *S*-facts.
> *Free lunch:* Containment relations are parsimonious.[14]

Summary. In this subsection, I have defined the central challenge for the physicalist position: to find a connection between physical ontology and other ontologies that does not carry an ontological cost. I have introduced the term *ontological supervenience* as a placeholder term for whatever the *for free* connection(s) might be. I then defined entailment as a kind of containment relation, observing that it could do the job that ontological supervenience needs to do and that it is a priori. Finally, I noted that the argument in chapter 2 was a direct argument that the physical facts could not a priori entail the facts of consciousness.

3.3 Appeals to A Posteriori Necessity

Are there kinds of ontological supervenience other than entailment? Many physicalists use an idea of metaphysical necessity that they take to be deeply tied to notions of a posteriori necessity. Philosophical appeals to a posteriori necessity appear in areas as diverse as the philosophy of mind (e.g., Levine 1993), the philosophy of causation (e.g., Fales 1990), and discussions of ethical realism (e.g., Brink 1991). These kinds of appeals raise the specific question of whether there is an a posteriori kind of necessity, such as metaphysical necessity, that can meet the restrictions on ontological supervenience.

Having dispensed in chapter 2 with the idea of an entailment from the physical facts to facts about consciousness, here I explore the different ways an a posteriori necessity might connect them instead. My method is to question the existence and nature of a posteriori necessity, exploring some ways of understanding its basis. In section 3.3.4, "The Minimal Meaning Postulate," I explain in detail what would count as a "basis." The short version is this: A *basis* for a posteriori necessity would be an intelligible constraint on the space of epistemic possibilities that excludes epistemically possible worlds from the resulting set of metaphysically possible worlds in a principled way.

Short of the full discussion in 3.3.4, there are three quick and overarching reasons for accepting the requirement that a posteriori necessities should have a basis. First is that the necessity of identity, which underlies many claims of a posteriori necessity, is an a priori principle. This general claim about identity is an intelligible basis for the necessities secured by it, and it is only the specific scientific identity claims that are a posteriori.

Second, on examination it is clear that the methodology for arguing that a posteriori necessities exist is the use of thought experiments. Because thought experiments give us only descriptions of situations about which we can make a priori judgments, the methodology itself shows that whatever basis we have for believing in a posteriori necessities is intelligible a priori.

Finally, we can observe that the logic of necessity is basically the logic of a universal quantification. For a supervenience conditional to be true necessarily, it must be true in all possible worlds. The logic of proving a universal quantification is that we can show that the conditional is true for *all* cases if, and only if, we can show it is true for *each* case. The fact that we should be able to show that the necessary conditional is true in *each* possible world suggests that in each world there is a basis for the connection between the facts such that, by understanding this basis, it is possible to show that the corresponding conditional will be true in that world.

3.3.1 Essentialism

Essentialism is the view that entities have certain of their parts or properties necessarily. An essence, as a necessary part of a thing, is something that thing has in

all possible worlds in which it exists: Presumably the essence of gold is the atom Au in the periodic table; the essence of light is the photon; and the essence of the atomic nucleus is the proton. Appeals to empirically discovered essences are one kind of appeal to a posteriori necessity. Thus a physicalist might propose that consciousness is essential to certain brain states in whatever a posteriori way empirical kinds come to have essences, and so the failure of entailment the antiphysicalist argues for is not fatal.

In *Identity and Necessity* (1971) and *Naming and Necessity* (1972) Saul Kripke provided a model for understanding a posteriori essentialism. Most philosophers appealing to a posteriori essentialism rely on Kripke's model. Around the same time, Hilary Putnam (1973) offered related arguments taken to yield a similar moral. This is an area in which there is much controversy, but it seems to me that Chalmers and Jackson have pointed out the basic problem with this appeal and that it was already present and discussed by Kripke in the footnotes to his own seminal work. In footnote 17 in *Identity and Necessity* (1971), Kripke himself alludes to the Chalmers-Jackson concept of *worlds viewed as though actual*, noting that it is distinct from the subjunctive point of view used to discover the kinds of possibilities that have come to be called metaphysical.[15]

A posteriori essentialism relies on rigid designation to guide how we should talk about worlds viewed as counterfactual. In Kripke's work, a rigid designator is a term that refers in a counterfactual world only if that world contains whatever the designator refers to in the actual world (if the term refers in the actual world). This means someone trying to use a rigid designator to describe another possible world is restricted in the description by actual world facts. Rigid designation provides the basis for a necessity knowable only a posteriori through this dependence on truths about the actual world to govern how statements are evaluated for truth in nonactual worlds.

However, because it is based on truths about the actual world, Kripke's essentialism helps the physicalist answer the antiphysicalist arguments only if entailments from base facts to higher-level facts are not required in the actual world, particularly as a way of connecting essences to their observable manifestations or securing facts about reference. The problem for an essentialist who makes this claim for essentialism is that rigid designation is silent about things such as how lower-level facts determine facts about reference and how essences produce their observable manifestations. Therefore, an appeal to rigid designation alone begs the question against the antiphysicalist, who maintains that entailments *are* needed in the actual world, at least by a physicalist who wishes to get such facts for free. To put the point in terms of the Kripke-Chalmers-Jackson distinction between ways of regarding worlds (as counterfactual or as actual), the subjunctive truths regarding nonactual worlds about which rigid designation provides guidance do not provide a basis for determining indicative truths about the actual world.

The challenge given to the physicalist is to produce a way besides entailment to get such facts for free, and an a posteriori essentialism that presumes and is therefore silent about the actual world facts provides no help. Rigid designation

is not the proper sort of thing to reveal or secure determination of phenomenal properties by physical properties in the actual world. As an analogy, "heat" is usually treated as a rigid designator, along with the claim that molecular motion is an a posteriori essence of heat. However, the necessary connection between molecular motion and the observable manifestations of heat is not secured by rigid designation, nor does rigid designation secure the fact that the reference of "heat" involves molecular motion in the actual world. Rather, rigid designation, *given those facts*, projects rules for the proper use of the term *heat* at counterfactual worlds. The determination of those facts from the physical facts in the actual world must be provided by something else.

Neither does the causal theory of reference solve this problem for the physicalist. The same issues can be raised about causal relations. If physicalism is true, causal relations between language users and high-level properties such as phenomenal properties must be determined by the basic physical facts. If so, the physical base of facts either entails them or it does not. The antiphysicalist arguments show that the physical base facts cannot entail facts about causal relations to consciousness (because it cannot entail facts about consciousness), which implies that there is some basis for a posteriori necessity being assumed rather than explained by an appeal to the causal theory of reference.

In summary, because the semantics of rigid designation treats the actual world facts as given, an appeal to a posteriori essentialism as a way of answering the antiphysicalist arguments requires making a simultaneous appeal to a basis for a posteriori necessity independent of rigid designation and responsible for physical facts determining the phenomenal facts in the actual world. It is this independent basis for a posteriori necessity that must do the work of answering the antiphysicalist arguments. Nothing in Kripke's work accounts for it.

A possible response to this interpretation of a posteriori essentialism is to complain that some conceivable worlds *must* be ruled out as possible by rigid designation because, for example, Kripke showed that a world in which H_2O is not the essence of water is conceivable but not possible. Similarly, it is natural to suggest that a world in which consciousness is not an essence of brain states might be conceivable but not really possible. Although this is a tempting analogy, I believe there is a fundamental disanalogy already pointed out, not only by Kripke but by others, also.

When considering the possibilities for what water might be, XYZ or H_20 for example, the fundamental ontological content of the rival possible worlds is represented by the nonrigid specifications of them as XYZ and H_2O worlds. Centrally, XYZ and H_2O worlds are supposed to be observationally indistinguishable, which implies that it is indispensable to the conceivability of the two worlds that the presence of either XYZ or H_2O (along with the appropriate contextual facts) can determine the manifestation of water's identifying characteristics. In these thought experiments, if there were some key features of water that H_2O or XYZ (along with the appropriate contextual facts) could not determine, then the deficient entity would not be a legitimate candidate essence for *water*.

It is precisely at this point that the parallel between the water case and consciousness seems to break down. Although the facts about H_2O do entail that the identifying macrocharacteristics of water will exist (if the appropriate context is provided), there is no such relation between the physical facts and facts about consciousness (and the case for lack of entailment was made in the last chapter without any essential appeal to the conceivability of a world). If a posteriori essentialism is to become relevant to the discussion about consciousness, there must be an independent necessity determining the higher-level facts, such as causal and phenomenal facts, from the lower-level facts, just as facts about H_2O and its context determine the observable facts about water.

The overall moral is that facts about a posteriori essences usually follow from conceptual necessity (entailment from lower-level entities to observable features of an explanatory target), some indexical truths (brute facts about what is present in our context), plus some facts about language (the rules of rigid designation). Before the Kripke and Putnam rules even become relevant, we must first decide on independent grounds what the situation is in the actual world. Does fire come from the release of phlogiston? Are the properties of water best explained by XYZ? Does cognitive neuroscience fully explain phenomenal consciousness? These questions are all on a par, and, most important, we settle questions about reference by *first* settling questions such as these. A posteriori essentialism comes later. Whatever independent evidential arguments there might be for physicalism, they must ultimately point to a *for free* determination relation from physical facts to phenomenal facts, and rigid designation is not up to the job.

Summary. Putting things another way, pointing out the failure of entailment is just a more philosophical way of pointing out that the physical facts alone fail to imply some observable facts about phenomenal consciousness.[16] Even after Kripke and Putnam, something cannot be an essence without that kind of entailment. Because this a posteriori essentialism implicitly requires appealing to a further basis for the determination relation before it can answer the antiphysicalist arguments, we cannot know if it meets the conditions on ontological supervenience until we know the basis of this further kind of determination relation. One commonly appealed-to candidate is the identity relation.

3.3.2 Empirical identity

The previous subsection argued that Kripke-Putnam's essentialism does not by itself provide an a posteriori yet *for free* connection between the physical facts and other kinds of facts. Many physicalists gain hope from closely related examples of empirical identity. For instance, the fine-grained fact expressed by *Cassius Clay was a great fighter* does not entail the fine-grained fact expressed by *Muhammad Ali was a great fighter*, but there is no possible world in which Cassius Clay is a great fighter and Muhammad Ali is not (since they are the same person).

The physicalist suggestion goes like this. Identities are necessary, so any possible world in which Cassius Clay fights greatly is a possible world in which Muhammad Ali fights greatly, and vice versa. The example shows a necessary connection between facts that seems to be based on empirical identity rather than entailment. Physicalists argue that examples such as this show that identity, a *for free* necessary connection, can exist without entailment.

Many people believe that conscious states are identical with special kinds of physical brain states (e.g., Papineau 1993). As mentioned earlier, views such as Christopher Hill's or Brian Loar's also seem committed to it. Even philosophers who share the antiphysicalist's views about the severity of the explanatory failings may hold this view. Joseph Levine, for example, has endorsed the existence of an "explanatory gap" between the physical facts and the facts about consciousness. He explores the explanatory gap in some depth (Levine 1998) but resists the antiphysicalist conclusion by suggesting that an identity holds. Levine proposes that identity connects brain states to consciousness necessarily and *for free* even though the facts about the former do not entail the facts about the latter. I call this kind of identity without an entailment from lower-level facts to higher-level facts *primitive identity*.

Methodological discussion. Here I present an informal, broad discussion of the principles broached and issues raised by the attempt to make a primitive identity claim. I chiefly discuss issues of method and the role identities play in justifying scientific explanation of an entity's properties. In a later subsection, titled "Ontological Discussion," I make a detailed and formal argument specifically that primitive identity cannot provide a basis for ontalogical necessity because, like rigid designation, primitive identity also requires an appeal to some further kind of a posteriori necessity.

If an empirical identity exists, then the identical objects must have all the same properties. I call this state of affairs *indiscernibility*. Let a *natural property* be a property that is causally involved in determining the dynamics of a spatiotemporal entity or that is determined by natural properties. Two entities are *naturally indiscernible* if they have all the same natural properties. Entities that are naturally indiscernible will have the same locations, the same masses, the same shapes, and the same internal structures. They will instantiate the same dynamics and enter into the same patterns of interaction with other entities. Any "two" entities that are identical will be naturally indiscernible.

The primitive identity theorist observes that identities are useful in establishing natural indiscernibility. Consider Joseph Levine, who, in his discussion of physicalism and identity, observes that we can derive the liquidity of water from chemical theory using the supposition that water = H_2O; he then claims that this supposition itself does not need to be derived. This is a difficult issue. I think the role that identities play in such derivations needs careful scrutiny before we can conclude that they might be helpful to physicalism, and I believe that they do not ultimately survive the scrutiny.

To begin, notice that in deriving the indiscernibility of water and H_2O, we use chemical theory along with other empirical identities to derive facts such as *batches of H_2O are liquid at room temperature*. We also use the identity *water = H_2O* as a bridge from the conclusion about H_2O to the same conclusion about water. We can verify that water has such a derived property, liquidity for example, using experiment or observation, adding support to the theory.

In the derivation, the identity is a bridge that transfers properties derivable of H_2O into hypotheses about water (tested by observation) or properties observable of water into hypotheses about H_2O (tested by derivation), but it does no *other* work. In particular, the identity cannot be used to introduce properties of H_2O other than those derivable from the theory. Any activity that uncovers nonderivable properties in the explanatory target instead results in reasons to modify or reject the theory. This restriction on the use of identity in scientific inference is critical both ontologically and epistemologically.

The restriction on the use of identity is critical ontologically because the inferences involving identity must respect the fact that it is the lower-level facts about molecules of H_2O that determine the properties of the higher-level entity, which is a volume of water. Because, using entailment, we can independently determine that the properties would be produced by the lower levels, it is harmless to use the identity to transfer these properties as provisional hypotheses about the higher-level entity. This procedure is consistent with the direction of determination.

But imagine if we allowed ourselves to use the identity to transfer properties the other way, attributing properties to volumes of H_2O that were not entailed by the theory, adding them solely because they are observed of water. If we did not then raise serious questions about the adequacy of the theory, it would raise serious questions about the direction of ontological determination. In the absence of the conclusion that some incompleteness has been uncovered in the lower-level theory, we would have no basis for preferring lower-level determination as the explanation for the presence of the property over some kind of strong emergence.[17]

The restriction on the use of identity is also critical epistemologically because it protects the coherence of scientific reasoning. In the primitive identity theorist's analogy, neurally characterized brain states are analogous to H_2O, a theoretical entity, and conscious experience is analogous to water, a commonsense observable. In the ideal case, the rational reconstruction of a successful identification would proceed by showing that:

- Water and H_2O (or brain states and conscious states) are naturally indiscernible.
- There is a base of facts entailing their indiscernibility, and this base of facts does not involve a circular appeal to the identity at issue. The base of facts consists of
  - the properties of the theoretical entity as entailed by the theory (supplemented noncircularly with provisional hypotheses about *other* empirical identities)

  - the properties of the commonsense entity as entailed by a set of observational facts
  - an exhaustive one-to-one mapping between the two sets of properties.
- Inference to the best explanation justifies our supposing an identity. The inference to the best explanation requires appealing to contingent principles such as simplicity, conservativeness, coherence, and the identity of indiscernibles.[18] Even so, it amounts only to adding an irreducible indexical fact about *our environment* (i.e., we live in an environment with H_2O *here)*. It does not justify supposing new *properties* had by any of the entities other than those that can be independently verified by theory (if they are theoretical entities) or observation (if the entity is an explanatory target).

The primitive identity theorist is essentially suggesting that we modify step 2 by using identity statements circularly to establish indiscernibility, giving up, even as an ideal, the standard of independence step 2 represents. In practice, the primitive identity statement would act as a bridge carrying properties both ways, from commonsense entity to theory and from theory to commonsense entity. For example, because there is an explanatory gap, the facts about brain states fail to entail some properties of experience, such as the experiencing of phenomenal qualities. Reacting to this, the primitive identity theorist proposes we use an identity statement as a bridge to transfer the needed subjective properties from consciousness to the brain states. Similarly, we must attribute to experience properties of brain states not observable in conscious experience, such as a fine-grained microphysical constitution.

In the normal case, whenever there is a need to carry properties primitively across the bridge, for example, from observations about water to our theory of H_2O, we would properly conclude that either the theory is deficient or our observations are deficient.[19] Metaphorically, we can say that whenever traffic crossing the identity bridge carries new properties along with it, we pay a toll: We acknowledge incompletenesses in the quality or ontology of the theory or deficiencies in our observation base.

Levine agrees with the antiphysicalist that the proposed base facts do not entail the natural indiscernibility between themselves and conscious experience, as the facts about the phenomenal qualities (at least) get left out. Levine recognizes this and so names it a *gappy* identity, but accepting the gappiness forces the identity bridge to carry traffic without collecting its toll, and the tollbooth plays a critical role. The standard practice of requiring the facts to entail indiscernibility *without* appealing circularly to the empirical identity at issue provides a mechanism for systematically testing and falsifying theories.

For example, we know that water is liquid at room temperature, expands when frozen, freezes at 32 degrees Fahrenheit, is transparent, dissolves salt, and so forth. To identify batches of H_2O with water, the facts about these batches must deliver a guarantee that all these water properties would be present if H_2O were

present. Entailment works nicely because it is a containment relation, showing that the theories convey the empirical information needed to deliver the guarantee. When entailment is absent, one can assume that some empirical information is missing from the theory.

By allowing entailment to be absent from empirical explanations of indiscernibility, primitive identities give incomplete or inaccurate theories an "out" for explanatory failure. A primitive identity would allow us to maintain, for example, that a theory of H_2O is complete and adequate even if it failed to entail the transparency of liquid batches of H_2O molecules. Allowing that kind of failure in the theory could undermine the credibility of the science, yet it is exactly analogous to the proposed use of primitive identity to explain consciousness. Our ideally completed physical neuroscience plays the role of a theory of H_2O, phenomenal properties play the role of the transparency of water, and the primitive identity theorist is like a hypothetical chemist who maintains that the theory is adequate despite not entailing transparency.

From ideal explanation to real explanation. At this point physicalists may object that the preceding account is unrealistically ideal, that we never do know if theory entails all the observable properties of our explanatory targets and nothing else. In general, we know that the theoretical entity and the explanatory target have some properties in common, we use inference to the best explanation to introduce the identity, and then we use the identity aggressively to derive other properties that they have in common, thus completing the case for indiscernibility. And, moreover, this is exactly what physicalists suggest we do for consciousness. We know that a conscious state and a brain state both produce certain behaviors (for example), so we use inference to the best explanation to suppose that they are identical and use the identity to derive their indiscernibility.

Admittedly, because we never really have *all* the facts, and because our deductive powers are limited, we often justify identity claims on grounds less compelling than discovery of entailments establishing natural indiscernibility. But these justifications always carry force because they rest on *evidence* of an entailment that eludes us because we are missing some facts or because the case is too complicated. The ordinary case, therefore, always leaves it open that there is or could be an entailment showing indiscernibility if we only had a more complete theory, more information, or greater powers of reason. We use the identity aggressively not to establish indiscernibility but to make hypotheses about potential entailments that we should require our theories to produce or for potential experiments whose outcomes would test the theory.

These ordinary cases are nothing like what we have with consciousness, for which the antiphysicalist has shown a clear in-principle failure of entailment. This clear demonstration of failure closes the door to the fallbacks that we are ignorant of some physical facts or lacking powers of reason. The primitive identity theorist has already agreed with the antiphysicalist's pessimistic conclusion, which is why they have proposed using identity to shore up the failure by primi-

tively ascribing properties to the different entities. Given the pessimistic conclusion, it becomes very difficult to see how we can maintain an analogy to ordinary practice.

Furthermore, I have defined entailment as a containment relation between empirical facts, where there may be different kinds of empirical information in the antecedent (i.e., the supervenience base) than in the consequent (i.e., the supervening facts).[20] As defined, entailment plays an ontological role by providing a for-free connection between ontologies, and its ontological role is not addressed directly by the pragmatics of practice. Even if we routinely make pragmatic decisions to assert and use empirical identities absent of knowing that there is a full entailment of indiscernibility, this does not imply that there is a different ontological ground for indiscernibility, allowing us to assert identity even when we know entailment fails. The antiphysicalists have produced forceful arguments that entailment is indeed missing, and the more formal argument against primitive identity that I develop subsequently is put in terms of the ontological significance of that conclusion rather than epistemology. Therefore, it is not directly addressed by observations of practice.

The foregoing implies that *if* there is an empirical identity between a collection of physical tokens *A* and a higher-level token *B* despite the absence of entailment from the *A*-facts to the *B*-facts, then good methodology forces us to conclude that we are missing some facts about *A* or *B*. With this in mind, I can now state the most important conclusion of this subsection:

Mind/brain identity is not a sufficient condition for the truth of physicalism.

All mind-brain identity implies is the existence of a monism, but not necessarily a monism of physical (or physicalistic) properties. Mind-brain identity is consistent with dual-aspect theories, dual-property theories, and other sorts of neutral substance monisms or nonsubstance (e.g., process-based) ontologies. Without an entailment from the physical facts to the facts of experience, one is merely allowed the conclusion that, if the mind and brain are identical, they are two aspects of something about which there are some further connecting facts to know.

Ontological discussion. The physicalist is proposing primitive identity to provide an a posteriori necessary connection between physical facts and facts of consciousness. I argue that *if* primitive identities exist, they cannot provide a basis for a posteriori necessity because primitive identities themselves would need to inherit their primitiveness from an independently based a posteriori necessity. The moral of my arguments relates to parts of Yablo (1987). Like Yablo, I argue that facts about token identity must supervene on other, more fundamental facts. The arguments here also relate to those in Johnston (1992). In outline, my argument is as follows:

1. Narrow facts about identity (e.g., identities with cognitive significance such as morning star = evening star) supervene on other kinds of narrow facts.

2. Either entailment provides the supervenience connection or it does not.
3. If entailment provides the connection, the identities are not primitive identities.
4. If entailment does not provide the connection, the primitiveness of the identities is inherited from the primitiveness of an underlying a posteriori supervenience relation.
5. Therefore, primitive identities result from rather than form the basis of a posteriori necessities.[21]

Primitive identity claims worry me because the truth of a high-level identity statement $a = b$ should supervene[22] on (1) facts about indiscernibility, most relevantly natural indiscernibility,[23] and (2) indexical facts about what lower-level entities are present. Informally, the analysis of the supervenience base for an identity is, "If two natural objects occupy the same spatiotemporal location and have precisely the same properties,[24] then they are identical." That's what I mean by saying high-level identities supervene on indexicality and indiscernibility.

For example, we think Venus, the morning star, and the evening star are the same heavenly body because we discovered reasons to believe that they have the same spatiotemporal coordinates and the same properties. If these things are true, they establish the identity.

Also, there is a basic deflationist intuition about identity underlying my worry. The deflationist intuition can be explained through an analogy with a deflationist position many philosophers have about truth. The deflationary position about truth is that asserting the statement *'p' is true*, for any proposition *p*, says no more and no less than assertion of the statement *p*. Deflationists about truth propose that the predicate 'true' marks a logical property of a sentence useful because it enables semantic ascent but does not assert a fact about a natural property of the sentence or its relation to the world.[25]

In an analogous way, the deflationist about identity proposes that the statement $x = x$ asserts nothing about x except a kind of logical property applicable to all things. Just as "truth" is useful for semantic ascent, empirical identity is useful because it allows us to organize and connect apparently disparate aspects of the world in the face of incomplete information. For example, if we do *not* have all the information about the morning star and learn that it is identical to the evening star, supposing the identity can help us consolidate and organize previously independent facts, and thereby to learn things about the morning star (and the world in general) that we otherwise might not be able to know.

To take another example: As Lois Lane learns scattered facts about Clark Kent and about Superman, by assuming their identity she may be able to fill in some cracks and make more sense of her knowledge than she otherwise could. For example, she would understand why Clark and Superman are never seen together and why Superboy always showed up in Smallville, where Clark grew up.

In each of these cases, the identity acts as a logical bridge enabling us to organize a partial set of facts, but it introduces no new natural facts into the world.

In general, there simply are no natural property facts to discover about x by knowing $x = x$ that are not discoverable from the other facts involved in individuating and investigating x.

When we move from considering trivial identities such as $x = x$ to informative identities such as $x = y$, where x and y are cognitively different referring terms, what changes is that the terms on either side of the identity relation may have their origins in different ways we have of knowing the referent. The significance of the identity relation as a purely logical property does not change. Because the identity, as a logical relation, still cannot introduce new natural properties to its referents, informative identity statements such as $x = y$ could not allow us to discover natural facts about either x or y that are not in principle discoverable by knowledge of x *as* x or y *as* y. If an identity such as $x = y$ is informative in a way that allows us to discover new things about x or y, it is only because there is some other information we are missing that accounts for the link between its different properties or aspects.

Consider again the example of Muhammad Ali and Cassius Clay fighting greatly. Imagine that Lois Lane, while taking a break from her investigation into Superman's identity, learns about Cassius Clay being a great fighter and having beaten Sonny Liston and also learns about Muhammad Ali being a great fighter and having beaten Joe Frazier. The identity between the two would be informative for her, but only because there are wide facts she does not know, such as that Cassius Clay changed his name to "Muhammad Ali" later in his career. However, all these wide facts are of a sort she could discover without using the identity, and the truth of the identity itself supervenes on these kinds of facts.

For clarity of illustration, assume that there are only two levels of facts, "higher-level" facts of the kind participants in the discussion about consciousness debate and "lower-level" facts of the kind found in the physical base or easily recognizable as constituted from them. It is hard to deny the supervenience of high-level identities. Facts about the *indiscernibility* of higher-level things will supervene on the base of lower-level facts, and the facts about higher-level identities are fixed once these facts about indiscernibility are fixed. There is no coherent way to suppose that a logical relation such as identity is involved in producing indiscernibility, so it is hard to see how facts about high-level identities could be prior in any important way to facts about indiscernibility.

The supervenience of identity makes it hard to see how primitive identities could be appropriately primitive. I now provide a more formal argument that the preceding points create problems for primitive identity. Assume that A designates an entity via a description of its lower-level constitution. The A-facts are facts about lower-level physical objects such as molecules, including facts about their properties, their spatiotemporal locations and relations, and their interactions.

Assume that B designates an entity via a higher-level description. The B-facts may include facts about things such as B's thoughts, feelings, and desires. These are higher-level entities whose nature and presence is not explicit in the base of A-facts.

Assume that some of these higher-level properties are *local properties*. A higher-level property is a local property if the entity has the property due to intrinsic facts about its own constitution. Two entities are *locally indiscernible* if they have all the same local properties. Facts about phenomenal properties are plausibly local.

Finally, *let* $A = B$ be a putative identity.[26] Here are two premises.

1. If $A = B$, then A and B are locally indiscernible.
2. The A-facts determine[27] the local higher-level properties possessed by A.[28]

Here is an argument that there is a problem with primitive identity. From premise (1) and the assumption of their identity, we know that A and B are locally indiscernible. By the law of excluded middle, the A-facts and B-facts together either entail that A is locally indiscernible from B or they do not. If the A-facts and B-facts do entail the local indiscernibility of A and B, then $A = B$ is not a primitive identity because the supervenience of identity tells us that $A = B$ would follow from the entailed indiscernibility plus indexical facts.

So we know that for $A = B$ to be a primitive identity, the A-facts and B-facts together must *not* entail the indiscernibility of A and B. If so, their local indiscernibility must follow from the A-facts and B-facts plus some other facts. The likely candidates for the further fact or facts are (1) a basis for necessity other than entailment or identity or (2) the identity itself.[29] Consider case (1): The primitive identity rests on an a posteriori necessity and is not providing a basis for it.

Consider case (2): The A-facts and B-facts alone do not entail the indiscernibility between A and B, so there is some local high-level property P such that the A-facts do not entail *P(A)* but the B-facts do entail *P(B)*.[30] From $A = B$, we can infer *P(A)*. By premise (2), we can infer that the A-facts determine (but do not entail) P. This determination relation, whatever it is, will be partially or wholly responsible for producing the indiscernibility of A and B because it is the relation that produces P from the A-facts. Because identity is a logical relation that does not produce new properties in nature, we know that identity is not the determination relation we are seeking. By the supervenience of high-level identities, the primitiveness of $A = B$ is therefore attributable (partially or wholly) to whatever determination relation is responsible for the presence of P. So again there is some independent basis for a posteriori necessity required for the primitive identity to obtain.

The preceding discussion creates difficult problems for a Loar/Hill kind of physicalist. Recall that the Loar/Hill physicalist holds that we apprehend consciousness under two psychologically independent classes of concepts with the result that simultaneously there is no entailment between the physical and phenomenal facts but there is a primitive identity. The methodological discussion shows why this is a very problematic assumption to build into a science, and the ontological discussion shows that it is an incomplete view, requiring some further basis for a posteriori necessity beyond the primitive identity.

Recall also the related views of a McGinn type of physicalist who holds that the conceptual distance between physical and phenomenal facts means the physical might entail the phenomenal but we can never see how. A physicalist sympathetic to a McGinn type of view might respond to the preceding argument using an opaque entailment strategy. For example, his or her response could be that *P* is after all entailed by the *A*-facts, but entailed in such a way that we cannot recognize it because the *A*-facts portray P under different aspects or modes of presentation than is available from the point of view that delivers the *B*-facts.

Essentially, a McGinn physicalist might claim that the *A*-facts entail all the local higher-level properties of *A* but we fail to recognize the entailment because of some related thing that is an obstacle to our recognition, such as differently apprehended aspects or modes of presentation of those properties that we do not have the conceptual ability to bridge. As a first response, it seems to me that if a local higher-level property has such-and-such-an-aspect or such-and-such-a-mode-of-presentation, then those things are themselves second order properties and their presence is addressed by the argument. However, even if they are not admitted to be proper properties, the preceding argument can be iterated as needed:[31] The *B*-facts entail that there is an aspect of *P*, Aspect(*P*), not entailed by the *A*-facts. But then Aspect(*P*) is determined by the *A*-facts without being entailed by them, and the conclusion is the same.

I believe a counter like this is always open for the anti-physicalist side. For example, if a physicalist reiterates the opaque entailment response by holding that the *A*-facts opaquely entail the property *aspects* because the aspects are conceived under different modes of presentation by different concepts, then the preceding argument, with minor variations can also be repeated for *modes of presentation of aspects of properties* and so forth. Eventually this variety of physicalist should identify whatever sort of fact (about a property, aspect, guise, mode of presentation, etc.) he or she thinks is responsible for the opaqueness of the *B*-facts relative to the *A*-facts. In the end, even on the opaque entailment position, it seems there should be *something* in the observation base of *B*-facts that is not entailed by the *A*-facts, and so, by parity of argument, we can show that making a primitive identity claim between A and B requires appealing to some unaccounted for form of a posteriori necessity.

The McGinn kind of physicalist does have it open to them to bite a bullet and claim there is nothing at all in the observational base of *B*-facts not entailed by the *A*-facts, but then they must suppose some kind of delusion undergirding the argument in chapter two. As I discussed earlier in section 3.2, I believe this final hard line puts the opaque entailment strategy on the slippery slope to Dennettian eliminativism and is more radical than we should accept.

Finally, a physicalist might respond that the difference between the *A*-facts and the *B*-facts is that the *B*-facts must contain an indexical fact that is essential to their entailing Aspect(*P*) and that this indexical cannot be reproduced within or added appropriately to the base of *A*-facts. However, reflection on the kind of gap we are trying to close makes this doubtful. We are trying to understand what we

would need to add to a base of bare differences for them to entail facts about phenomenal quality. Even if it is true that facts about phenomenal quality contain essential indexicals (and I think that is far from obvious, as it is natural to presume that many different people from many points of view can experience similar or identical qualities), there seems no reason to suppose that an indexical fact *alone* can bring a world of phenomenal qualities out of a world of bare difference. Before accepting that an indexical fact could have that kind of precise significance, we would need a very substantial theory of indexicals that we do not currently have.

Summary. If the preceding analysis is correct, to give up entailment and still maintain rational coherence a theory of primitive (or gappy) identities must appeal to some basis for a posteriori necessity that will guarantee the natural indiscernibility of the identical entities, and this leaves it with the challenge of explaining the basis of this necessity. Thus it becomes clear that primitive identities cannot really be the basis for a posteriori necessity because the primitive identity claim is actually assuming the existence of a basis for a posteriori necessity. Therefore, primitive identities meet the restrictions on ontological supervenience only if this further a posteriori necessity they implicitly require meets those restrictions. A common suggestion for this further necessity is the necessity of natures.

3.3.3 Necessity of natures

Physical things have natures of their own existing independently of our concepts and their meaning. Having seen that a posteriori essentialism and primitive identities do not provide appropriate connections between physical and phenomenal facts, necessities that are in the nature of things are the next place to look. Perhaps the for-free connection between the physical facts and the facts of experience has its basis in the necessity of natures. In fact, primitive identity theorists, who I have argued must take the primitiveness of primitive identity as coming from a primitive indiscernibility, might naturally look to the necessity of natures to produce that primitive indiscernibility. Furthermore, our knowledge of the nature of things is empirical, so we should expect the necessities of nature to be a posteriori. If so, this would seem to answer the antiphysicalist arguments, which physicalists hold only target a priori necessity.

The natural sciences are full of examples of necessities due to natures. Returning to my canonical example, science postulates that water has a nature, H_2O. The conceptual content of the theoretical term H_2O represents that nature, and its theoretical context fixes its content. Within that context an assumption that something is H_2O entails certain consequences and is consistent only with a restricted range of facts. Entailments represent the necessities due to the nature of H_2O, and consistencies are the possibilities for that nature.

For example, the conceptual content of the theoretical concept H_2O entails that anything that is H_20 is liquid at room temperature.[32] Finding the entailment requires making suppositions about the shapes of the molecules and deriving facts about the orbital shells surrounding them, how freely electrons can move within them, how this affects the cohesion between the molecules, and how this affects the behavior of aggregates of such molecules at room temperature. Finally, one must do a bit of conceptual analysis on what it means to be *liquid* to see if such behavior realizes a liquid state (i.e., makes it indiscernible from a liquid). This entailment is a necessity from the nature of things; and, the antiphysicalist arguments demonstrate exactly that no such entailments exist between the physical facts and laws and the facts of experience.

One other place necessity shows up is in the derivation of behaviors that the electrons will exhibit, such as their freedom of movement between orbitals. To derive electron behavior in the depicted circumstances, one appeals directly to axioms of the theory. The shapes of orbitals and electrons' orbital behavior are derived from interference effects between possible paths the electrons take around the nucleus. Unlike entailments, these axioms do represent a posteriori necessities, but these necessities are natural laws. The laws do not constitute *for free* connections (as discussed in section 3.2).

What about cases in which we accept that there is a necessity even without an entailment from theory to phenomena? Examples like this abound in cosmology, for example, where cosmologists argue about why the universe is so clumpy at a large scale despite its lacking enough visible mass to account for its clumpiness. Examples like this do not help, because science treats the related theories as incomplete. A theory's lack of explanatory power, understood as its inability to entail certain facts, is the only real kind of evidence one could have that it is incomplete.

Physicalists need an example of a theory that the experts believe to be complete (or at least complete in the relevant respects), that fails to entail certain facts about its subject matter, and yet one in which it is acceptable to suppose that those facts follow necessarily from the natures of the theoretical entities described by the theory. It is implausible that such an example *should* exist, because that kind of predictive failure is always taken as compelling evidence that a theory is incomplete.

If a theoretical conception is accurate and complete, we should expect it to represent the necessities in its referent's nature, because to understand a successful conceptualization is to have information about its subject matter. A commitment to the information-bearing character of concepts is the most simple and direct way to make sense of the undeniable fact that deriving consequences from a theory is useful in the first place.

The view that understanding a conception is to have information about its subject matter provides a simple answer to the question: Why should a derivation of the logical consequences of a *theory* enable us to discover new facts in the

world? When we engage in this activity, we are leveraging the information *explicitly* contained in our conceptualization of the relevant entity to discover information, and therefore facts, about our world that are *implicit* in the conceptualization. If the concepts did not bear information in the first place, the activity would be hopeless: garbage in, garbage out.

Thus a complete and accurate *conception* of a subject matter will contain complete and accurate *information* about its subject matter. The antiphysicalist arguments establish that a complete and accurate physical theory will not entail the facts of experience. It follows that the physical nature of things does not necessitate those facts. No room is left over for an a posteriori necessity of natures to do any work. With the failure of the necessity of natures to provide a basis for a posteriori necessity that is helpful to physicalism, we appear to be out of candidates. How does this affect the prospects that physicalism might be true?

3.3.4 The minimal meaning postulate

I have argued previously that essentialism, identity, and the necessity of natures all fail to provide a basis for an a posteriori necessity able to help physicalism. If these arguments are correct, we are no better off than we were when we started wondering what could be the basis of an a posteriori necessity capable of serving physicalist ends. Such a necessity could exist only if there was a world that seemed possible a priori but that was not really possible for reasons relevant to the physicalist/antiphysicalist debate. In this subsection I discuss what it means to wonder if there is a metaphysical grade of possibility beyond natural possibility but short of conceptual possibility. And I answer that it is not clear and that this lack of clarity presents a significant problem for physicalism.

To begin the argument, I suggest some very mild constraints that any kind of possibility statement must meet to be meaningful. Consider statements of the form *X is possible* and *X is not possible* as they occur under normal circumstances. Reflection on ordinary examples strongly suggests that a meaningful possibility (or necessity) statement must meet certain minimal standards. I can formulate these standards using three criteria:[33]

1. We always make possibility statements relative to an established or assumed context.
2. Understanding such assertions tacitly requires holding the truth of the context, or some crucial elements of the context, constant as a constraint on the claim.
3. An assertion of a possibility involves an assertion that the hypothetical situation in question can be part of a consistent extension of that context, or of whatever part of the context the speaker(s) is holding constant.

People use possibility statements widely in everyday life, and this large variety of possibility statements all seem to meet the preceding criteria. Consider a chess

player mulling over his options. A friend may suggest a move where, in fact, another piece blocks the path. In deciding that the move is not possible, he clearly is not deciding that the move is contrary to the laws of nature or logic. He is judging the more pertinent question of whether, in a chess game, the move would be consistent with the rules of the game.

Similarly, a worker in a complicated bureaucracy may inform customers of possible avenues they can pursue to have a complaint processed. Why is it that filling out a form for review by the manager is a possible course of action but storming the CEO's office is not? The possibility of the former action, but not the latter, rests in its consistency with the normal processes and rules of the organization.

For a possibility statement to be meaningful, the constraints do not have to be sharp or explicitly understood. In ordinary use, we might say that climbing Mount Everest would be something that I could not possibly accomplish. Someone who makes that claim does not really mean a contradiction is involved in the description or even that it would violate the laws of nature. Instead, that person is appealing to vague and implicit constraints on what common sense would allow that I could reasonably or ordinarily accomplish. To claim that something like climbing Mount Everest would be impossible for me is to claim that my accomplishing it would be inconsistent with those implicitly understood standards.

Relative to a context *C*, the intelligibility of *X is possible* seems to require at least that the truth of *X* would not violate the constraints implicitly taken from *C*. This seems encodable into something like the following *minimal meaning standard*. For a statement about possibility, the minimal meaning standard is simply:

> Definition 3.2: A modal statement *X is possible* uttered in a context *C* that contains background constraints *BC* meets the minimal meaning standard if, and only if, *X* is logically consistent with *BC*.

No statement that is missing background conditions capable of supplying the consistency constraint can meet the minimal meaning standard. I am not putting forward the minimal meaning standard as anything more than its name suggests. Specifically, it is not intended as an adequate analysis of modality. *MM* is just a requirement on the meaningfulness of particular modal statements or claims.

In the search for truly metaphysical constraints to underwrite metaphysical necessity and possibility, it seems that the relevant constraints are not being even obliquely specified, and this raises worries about whether any really exist. As a point of fact, many philosophers who might have appealed to Kripkean cases to explain what they mean by *metaphysical necessity* have recently given up those cases as analogous to the case of experience, often due to critiques by antiphysicalists such as Chalmers and Jackson. The problem they now face is that, without these cases to provide a constraint to satisfy the minimal meaning standard, they move into an area in which it is unclear that there is a notion at all. This

problem as yet seems to be little appreciated. The literature defending physicalism is still full of confident-sounding appeals to "metaphysical" necessity. I worry that everyone does not really know what everyone else is talking about.

The need to satisfy the minimal meaning standard leads to a very general problem for any physicalist who wishes to use a posteriori metaphysical necessity to save physicalism. If identities do not form its basis, if natural laws do not form its basis, and if Kripke-Putnam essentialism does not form its basis, it seems as if none of the purely naturalistic constraints can do the job. If there are some further constraints on the space of possibilities over and above the laws of nature, empirical identity, and the rules of language, it seems that they will have to be something like a posteriori laws of metaphysics, whatever these might be, and however we are supposed to discover them.

These laws of metaphysics will have to complement the laws of nature in some way, producing a more liberal constraint on the space of possibility than natural law but less liberal than conceivability. It is at this point unclear whether the idea makes sense, but even if it does, these extra "laws of metaphysics" will not yield a for-free connection the way entailment does. In our fundamental ontology, we will have to postulate all the physical facts and natural laws and, in addition to this, some set of *metaphysical laws* or *metaphysical constraints* that rule out such things as the Zombie world. So I think physicalism would be false all the same, and a posteriori necessity cannot save it, even in principle.

3.4 Appeals to Meaning Holism

The previous chapter argued that no entailment connects the physical facts to the facts of consciousness, and the last section of this chapter argued that an a posteriori necessity cannot salvage physicalism. Some physicalists reject the distinction between the a priori and the a posteriori completely, complaining that it illegitimately relies on the analytic/synthetic distinction. Instead, these physicalists embrace some form of holism, either of meaning or confirmation.

Their claim is that the antiphysicalist arguments fail within holist frameworks that reject the distinction between conceptual connections and empirical connections. They may even reject the ideas of propositions or information, and these are central to the concept of entailment I used. In this section I address meaning holism, arguing that the antiphysicalist arguments can be adapted to a holist framework. The entailments those arguments appeal to are no more suspect than any others in use inside science, and the explanation and ontology of consciousness remains problematic even within a holist framework.

Quine's rejection of the analytic/synthetic distinction. W. V. O. Quine gave birth to modern holism by rejecting the analytic/synthetic distinction in *The Two Dogmas of Empiricism* (1963). Analytic statements are those whose truth we can determine by knowing the meanings of their terms, and synthetic statements

are those whose truth requires referring to matters of contingent fact in the world.

The classic example of an analytic statement is, "Bachelors are unmarried." A less classic example is "The team that finishes with the best record in the National League East will finish in first place." These sentences are analytic because anyone who understands the words they are composed from can determine their truth without having to check a textbook or a newspaper.

Examples of synthetic statements are "Trey is having girl troubles," and "The Braves beat the Expos today." They are synthetic because, even if you know what they mean, you still have to know something further about the world before you know if they are true or not.

In *Two Dogmas*, Quine asked how we are to understand what is meant by *analytic* in the claim that

(1) No bachelor is married

is analytically true. Following Frege, Quine proposed that it was because

(2) No unmarried man is married

is logically true, and statement (1) is gained from (2) by substitution of synonymous terms. Having diagnosed the first class of these analytical statements in terms of logical truth plus synonymy of terms, Quine poses the problem of how to analyze synonymy. According to Quine, "synonymy is in no less need of clarification than analyticity itself." He considers a variety of proposals for understanding analyticity besides synonymy, including verification criteria and semantical rules. For modern readers, though, the force of his arguments rests on the failure to account for synonymy. For example, he rejects the idea that we can analyze synonymy in terms of definition because the relevant notion of definition presupposes synonymy. Quine fails to find an analysis of what synonymy of *terms* is and therefore fails to find an analysis of the analyticity of certain *statements*, taken in isolation. He concludes that the idea of analyticity is too murky to trust.

Based on these problems, Quine rejects the notion of analyticity of statements. Quine's efforts fall short of showing that no satisfactory account of analyticity exists, but he feels he has shown that we have no good reason to *believe* such an account exists. Consequently, those who continue to believe in it are being dogmatic. Those moved by Quine's worries have not found subsequent attempts to articulate the distinction to be satisfying, and skepticism about the distinction remains.

Traditionally, empiricists believed that we express a posteriori knowledge using terms and statements with *synthetic* meanings and that we express a priori knowledge using statements with *analytic* meanings (rationalists differ from empiricists by holding that some synthetic statements express a priori knowledge). The following table shows the cross connections.

	A Priori Knowledge	*A Posteriori Knowledge*
Analytic Statements	*Traditional empiricist daim* There are analytic statements that express a priori knowledge. *Quinean argument*: There is no reason to believe in analytic statements and so no reason to believe that there are statements expressing a priori knowledge.	*Traditional empiricist diam* Analytic statements do not express a posteriori knowedge. *Quinean argument*: All statements express knowledge taken from both experience and linguistic convention to some degree. The sharp distinction between a priori and a posteriori knowledge is not justified.
Synthetic statements	*Traditional rationalist daim* Synthetic statements sometimes express a priori knowledge. *Quinean argument*: There is no reason to believe in synthetic statements because we understand them by contrast to the dubious notion of an analytic statement.	*Traditional daim* (*agreed by all*): Synthetic statements can express a posteriori knowedge. *Quinean argument*: All statements express knowledge taken from experience and linguistic convention to some degree. The hard distinction between a priori and a posteriori knowledge is not justified.

Quine's arguments attempted to undermine the entire theory behind the analytic/synthetic distinction, and with them the distinction between a posteriori and a priori knowledge. In doing so, Quine left empiricists with the stumbling block of explaining how we can gain all our knowledge about the world from experience if we cannot even derive the meanings of ordinary terms from experience or produce a class of experiences that circumscribe truth-conditions.

Quine himself was an empiricist, and he needed to reorient empiricism by connecting knowledge exclusively to experience, without supposing that these connections involve analytic reductions of individual statements or leaving room for statements whose truth can be determined from their meaning alone. His solution was to turn to radical holism.[34] Here are a few quotes from the *Two Dogmas* in which he gives his solution [all emphases added]:

> *My countersuggestion, issuing essentially from Carnap's doctrine of the physical world in the* Aufbau*, is that our statements about the* external *world face the tribunal of experience not individually but only as a corporate body.*
>
> *My present suggestion is that it is nonsense, and the root of much nonsense, to speak of a linguistic component and a factual component in the truth of any* individual *statement. Taken* collectively*, science has its double dependence upon language and experience.*

> . . . *total science is like a field of force whose boundary conditions are* experience. *A conflict with experience at the periphery occasions readjustments in the* interior *of the field.*

As Frege had earlier rejected the idea that individual terms were meaningful in isolation from statements, Quine's radical move was to reject the idea that statements are meaningful in isolation from theory. Even theory, Quine suggested, was not meaningful except through the way that it is embedded in the whole of language. Thus the meaning of every term and every statement implicitly relies on its relations to every other term and statement in the language, and, ultimately, it is language as a whole that is responsive to experience.

Preliminary rebuttal. A tremendous amount can be and has been said about Quine's views, and I can only sketch the outline of an answer here. But before beginning my main critique of the meaning holist's objections, I need to address a related objection that also comes from Quine.

The objection comes from Quine's quasi-behaviorism and is directed at the way the antiphysicalist uses the term *experience*. Antiphysicalists cannot mean what Quine takes it to mean (e.g., Quine 1992): stimulation of the sensory nerve endings. That is not what a theory of consciousness will be about. *Experience*, as is often pointed out, is ambiguous and also refers to phenomenal experience, which involves the first-person experiencing of certain kinds of qualities.

This presents the holists with a dilemma. Either they insist, with Quine, that "experience" on "the periphery" is univocal, referring behavioristically to "stimulations of sensory nerve endings," or they admit to the legitimacy of experience, phenomenally construed. It is easy to see that taking the first horn of the dilemma does not help the holists. Hung on that horn, they are objecting to a premise that is supported by a strong auxiliary argument that qualia are observables, an argument given in section 2.4 of chapter 2. Horn one is really a bald-faced eliminativism and lacks plausibility.

The second horn is different. The qualities of experience show up as close to the periphery of experience as possible. Phenomenal experience resides in something like Searle's Background (1983),[35] waiting for science to hook up with it. More strongly, its instances *pervade* The Background. Various phenomenal concepts are *so* ubiquitous, our cognitive system seems close to hardwired in its stubborn insistence on throwing the phenomenal properties into the world by latching them onto our intensional objects: that feeling is *in my foot*; that color is *on the wall*; *the rose* smells so sweet. The qualities of phenomenal experience are everywhere, literally fused into our model of the world as intensional support for our understanding of ordinary objects. Any systematic failure of theory, as a whole, to account for the phenomenal qualities of experience constitutes a massive failure of the *internal* part of Quine's theoretical field (physical science) to account for pervasive elements on the *periphery* (roughly, The Background). By Quine's own lights, it is the internal elements of the field that must buckle.

I remove this element from Quine's philosophy. If we also reject his exten-

sionalism, the remainder is an intensional holism in which phenomenal consciousness exists as a central target of explanation on the periphery. I argue that this remainder is quite harmless to the antiphysicalist arguments.

The stability of conditionals. The antiphysicalist entailments are analytic *in some sense*, but not in a way that violates holist restrictions. They evade sin because they do not rely on judging the truth of single statements independently of theory or on accommodating single contrary experiences to theory. They are explicitly concerned with systematizing the relation of physical theory *as a whole* to phenomenal experience *as a whole*, avoiding trafficking in isolated statements of theory or comparisons to isolated experiences.

Some people might be uncomfortable even with the idea of entailments from theory as a whole to various sets of facts as a whole. These conditionals from the whole theory to those facts might be seen as analytic themselves. But nothing in Quine's arguments forbids entailments from theories as a whole to consequences of that theory, so long as the larger embedding network for that theory is held relatively constant.

For example, Euclid's axioms continue to entail the theorems of Euclidean geometry, despite the empirical preeminence of non-Euclidean theories of space. These entailments follow from the set of axioms as a group, and nothing in the larger web has caused us to reject the rules of inference or interpretations on which they rely. We simply reject the idea that they accurately represent the geometry of space, but that kind of empirical failure very rarely, if ever, affects the truth of the purely conditional conclusions at issue.

Antiphysicalists can reconstruct their arguments within a holist framework, and, within that framework, they have the form of demonstrations that a certain class of theories fail to *predict* certain facts about experience: the existence, character, and experiencing of phenomenal qualities. The antiphysicalist arguments are special instances of ordinary scientific falsification.

The role of systematization. Some people might object that the antiphysicalists cannot show failure of prediction without appealing to the specifics of a particular theory. But the antiphysicalist arguments *systematize* the failure of physical theory as a type, demonstrating the in-principle obstacles of an entire class of theories. Such in-principle arguments appear in scientific discussion elsewhere, and we do not consider them suspect because of Quine's worries.

Although necessarily less formal, these arguments have a form similar to that used by John Bell in showing that the predictions of quantum mechanics (QM) could not be reproduced by any local hidden variable theory. In making his case, Bell did not examine the details of every single possible local theory, demonstrating for each how it failed. Instead, he abstracted out the general conditions that limit any such theory and showed how no theory meeting those conditions could predict the same results as standard QM. In doing this, Bell showed how we could determine that local theories, in principle, could not do justice to expe-

rience on the periphery, conditional on standard QM being correct in its predictions of experience.

The antiphysicalist arguments make the same kind of point about the relation of purely physical theories to conscious experience, using the same method Bell used. They point to certain features shared by an extraordinarily broad class of theories and argue that a theory with just those kinds of features, and no others, will fail to predict some facts about experience. We can possess all the physical facts, without being able to derive the truths about the phenomenal facts. If this is true, then all the postulates of physical theory cannot do justice to some facts about experience on the periphery. The facts concerning its phenomenal character are absent. The kinds of entailments involved are not of a different type than science usually uses.

If one wishes to manufacture an epitaph for these entailments by calling them analytic, so much the worse for Quine's arguments against analyticity. Nothing about meaning holism invalidates Bell's Theorem. Then, prima facie, nothing about it invalidates the antiphysicalist arguments either, in form or content, because the practice of extracting predictive content from theories is more secure than the arguments advanced against analyticity. Perhaps the antiphysicalist arguments violate Quinean strictures in some subtle way, but meaning holism provides no easy and obvious in-principle objections.

Experience and The Background. If the previous discussion is correct, then entailment claims are "true by virtue of the meanings of the terms," but in an entirely innocent sense. It is the exact same sense featured in entailments from physical theory to experience generally: Given an understanding of the meanings of the terms, one can in principle determine the truth of the conditionals connecting the theory in the interior to the facts on the periphery. Every *successful* physical theory supports similar entailments. Quine's arguments show nothing more than how these entailments do not exist from statements within the theory, taken in isolation. Where they do exist is from the theory as a whole to elements of the periphery where we live and experience. These elements of the periphery do not have to be taken as given, incorrigibly, but they are rarely if ever definable, and they do not get their meanings from the science being tested as much as from what Searle calls The Background.

The Background is that vast, unspeakably subtle and intuitively understood network of acquaintance and conceptualization that is built up from living where our concepts are working; concepts arising from The Background are developed from our being in the world. Our grasp of the concepts that connect us to The Background has never before been so poor that we could not judge the requisite entailments that connect theory to it, and the antiphysicalist can see no special reason to think that with consciousness, sui generis, we should begin doubting.

By rejecting materialism, the antiphysicalist is simply following Quine's good advice: Science is a continuation of common sense, and it continues the com-

monsense expedient of swelling ontology to simplify theory. Although application of his advice in this case might earn Quine's consternation, the fact that he himself might not like it does not undermine its rational foundations. If good arguments exist for the position, as they seem to, we are left with no other choice.

Finally, meaning holism often goes hand in hand with the program of naturalizing epistemology, and the holist is often bothered that the antiphysicalist project is at odds with epistemology naturalized. By the end of this book, it will be abundantly clear that the question of what nature is like is exactly what is at issue here. One cannot judge whether a theory is compatible with naturalized epistemology without implicitly or explicitly appealing to some view of nature. Because Liberal Naturalism is challenging our view of nature, these objections beg the most crucial questions and are entirely moot. The meaning holist does not seem to have an objection that carries through.

Summary. The meaning holist objects that the antiphysicalists require the existence of analytic truths connecting the physical facts to the facts of experience. Although this is one way to formulate the arguments, they are also compatible with a holist framework for meaning once the extensional stance is rejected. All the argument needs is a distinction between concepts of experience existing ubiquitously on the periphery of the web of meaning, theoretical postulates about the nature of the physical embedded within the web, and a failure of inferential connection using standards of competence that we accept elsewhere and have no good reason to reject.

3.5 Appeals to the Danger of Causal Irrelevance

In no place have my arguments appealed in an important way to the logical possibility of zombies, physical duplicates of normal human beings who nevertheless lack conscious experience. In fact, the one standard antiphysicalist argument that I called on was Jackson's "Mary" argument, and even it played a weak role. I used it only to establish the observational fact that phenomenal qualities are not patterns of bare difference, by itself a far weaker and more easily defensible conclusion than Jackson himself drew from it.

Physicalists will still be worried. At the end of the day, if consciousness does not ontologically supervene on the physical facts, then it *does* seem as if zombies are possible. The possibility of zombies raises serious worries about the causal relevance of consciousness in the actual world and about the epistemology of consciousness generally. Sidney Shoemaker (1999), for instance, argues that rationality itself will be undermined if zombies are possible.

My own position is a middle one. Zombies frighten me less than they frighten many physicalists, but I am also less sanguine about the prospect of zombies than are many antiphysicalists. If zombies are not possible, and I believe they might not be, it will be because of extraphysical constraints on the space of possibili-

ties, so physicalism is still false. The possibility or impossibility of zombies plays no essential role in the truth of physicalism. Nevertheless, it is easy to see why the idea worries physicalists. They believe the only sure way to avoid zombies is to argue that, despite first appearances, the physical facts *do* entail the facts of experience.

Once the worry about zombies has taken hold, the final battle line for physicalism is to argue that the physical facts simply *must* entail facts about consciousness; otherwise, we open the door to deadly absurdities. The most promising way to argue the point is to appeal to the belief that consciousness must perform some function, and we have every reason to believe that it is the physical stuff of our brains that is performing the relevant functions. This view is very widely held.

In part II of the book I pursue what I think is a deep and interesting response to concerns like this. I think the central intuitions behind these arguments may be correct, at least in support of the weaker point that consciousness, where it exists, is invariant with respect to the conscious system's functional organization. Still, these intuitions do not yield the conclusion that consciousness is ontologically supervenient on the physical facts. According to the view of causation I develop, the antiphysicalists have room to evade the charge that consciousness must be epiphenomenal on their view, and they can do so without falling into an interactionist position.

I argue that physicalists tend to fall into the trap of overlooking certain subtleties involved in deciding whether two systems are functionally identical. Two systems are functionally identical just in case they are causally isomorphic at the appropriate level of organization, meaning that they have the very same causal organization at that level. Hidden in such judgments are some subtle questions about causation, levels of nature, and causal organization at a level. In "Faces of Causation," part II of this book, I argue that the facts about causal organization do not ontologically supervene on the physical, either.

According to the Theory of Natural Individuals developed there, causation itself has multiple aspects. The facts about some of these aspects are not strictly implied by the explanations of physical science or the story about functional organization that we would tell from a purely physical perspective. Whether or not a functional identity exists between two systems cannot be decided independently of knowing the totality of causal facts, and these facts are not just the physical facts. So it is possible to maintain that the facts about consciousness are invariant with the functional facts about a system but are *not* determined by the physical facts about the system.

More strongly, I argue that the physical facts *necessarily* underdetermine the functional facts. It is not a matter of there being some contingent nonphysical entity interacting with the physical. The nonphysical aspects of causation that I introduce do not interact with the physical in any way. They complete the metaphysical story about what interactions are in the first place.

3.6 Conclusion

Physicalist responses to the antiphysicalist arguments rely on making relevant distinctions between ways of knowing the world and the truth about its ontology. These replies cannot work if physicalism is true because a complete physical theory should give us complete knowledge, at least in principle, of the physical facts. We can see that attempts to use a posteriori necessities to bridge the gap between physical facts and facts of consciousness fail once we inquire closely as to the basis of the a posteriori necessity because these appeals actually imply that we are missing some facts. Furthermore, the antiphysicalists can adapt their arguments to holist frameworks.

4

The Boundary Problem for Experiencing Subjects

4.1 First Steps

If physicalism is false, we must look for an alternative way to place conscious experience in the universe. The alternative I explore is not a Cartesian dualism but a version of Liberal Naturalism. Liberal Naturalism is the view that nature is built on a single fundamental kind, and, if so, that some aspects or properties of this fundamental natural kind are not physical. Liberal Naturalists cast the problem of consciousness differently than do some who claim that it is the problem of reconciling consciousness with materialism. William Lycan (1996) expresses this view of the problem:

> It has to do with the internal or subjective character of experience, paradigmatically sensory experience, and how such a thing can be accommodated in, or even tolerated by, a materialist theory of the mind. (p. 1)

Lycan's statement of the problem makes a pretheoretical commitment to the salvation of a certain metaphysic, physicalism. To the extent that Lycan is representative, one could say that the physicalist's overriding priority is to be ontologically conservative, and that to honor this commitment, he or she has to pay the price of being methodologically radical. As chapter 3 discussed, their prior commitment to physicalism forces physicalists to take such measures as blaming theoretical failures on cognitive deficits of the theory makers rather than on the quality of the theories; approving of appeals to unique and not clearly meaningful kinds of necessity; postulating primitive identities; or arguing for the elimination of self-evident observables.

In contrast, Liberal Naturalism primarily wants to explain consciousness clearly, without appealing to anomalous standards of explanation. The Liberal Naturalist point of view is that the scientific enterprise accepts the discovery of natural ontology as its purpose. The thing keeping Liberal Naturalism honest is

not its commitment to a metaphysic but its rigorous standards for rational explanation. For Liberal Naturalism, setting aside these standards to save an ontological viewpoint is an unwise perversion of science.

Liberal Naturalism believes just this: The problem of consciousness is to understand why it exists; what its relations are to the other things we know exist; and what difference it makes, if any, to the natural order of things. Liberal Naturalism has weaker metaphysical commitments than physicalism because its primary allegiance is to the empirical project of explanation. One might suggest that Liberal Naturalism is metaphysics in the service of explanation, whereas physicalism is explanation in service to metaphysics. Accordingly, the Liberal Naturalist is methodologically conservative, and this conservatism will lead its adherents to be ontologically radical.

My form of Liberal Naturalism is a variant on a kind of view put forward before by authors such as Whitehead (1929), Russell (1927), Maxwell (1979), Lockwood (1989), Griffin (1998), and Sprigge (1994), tentatively endorsed by Chalmers (1995), and recently suggested again by Galen Strawson (1999) and Thomas Nagel (1998):

> It may be that the physical description of the brain states associated with consciousness is an incomplete account of their essence - that it is merely the outside view of what we recognize from within as conscious experience. If anything like that is true, then our present conceptions of mind and body are radically inadequate to the reality, and do not provide us with adequate tools for a priori reasoning about them. (Nagel, 1998)

So this suggestion arose here and there in the twentieth century. Russell suggested that the problem stems from science portraying matter structurally, focusing on its form and not its content. Restating Russell, Lockwood adopts some physicalist terminology from J. J. C. Smart and explains that our concepts of the physical are topic neutral, meaning that they say nothing about the basis of physical being. That is, physical theory says nothing about what physical things are like "in themselves." Whitehead called belief in the adequacy of such descriptions the "fallacy of misplaced concreteness" and argued against any such notion of vacuous actuality.

One can see a commitment to traditional materialism in the names that proponents of the view often give it. Maxwell called it Nonmaterialist Physicalism. Lockwood distinguishes between Physicalism, which he thinks is false, and Materialism, which he thinks is true. David Ray Griffin writes of Panexperientialist Physicalism. Galen Strawson calls it Realistic Materialism.

Many of these authors seem unwilling to go beyond physicalism in any way more radical than the hypothesis that the physical has an "inner aspect" tied somehow to experience. Seeing no more significance than this in the view will block efforts to fully develop it and make it viable. As I argue in the rest of the book, making this sort of Liberal Naturalism work requires undertaking more thorough revisions to our view of nature than these authors entertain.

Although Liberal Naturalism might feel liberating, we have too much freedom. To find a place for consciousness, we need tests for the minimal adequacy of proposed explanations and also a class of problems able to provide clues that help us triangulate to the point of fundamental incompleteness in our knowledge. As a beginning for the effort, I wish to step back to examine assumptions and to try to identify the deepest problems and clues in the vicinity.

Because we are searching for new facts about nature that are fundamental, the most helpful kinds of puzzles to focus on may not be specifically cognitive puzzles. By *cognitive puzzles* I have in mind the sorts of questions raised by facts such as: conscious states tend to be reliably reportable, conscious states are representational, conscious states contain information that is globally available in the control of behavior, and the fact that the structure of consciousness mirrors the structure of cognitive processing. All of those facts will be very important *eventually*, and any theory must allow us to understand why consciousness has those features. In the context of a foundational search, however, they are not likely to be the best pointers to follow. The next few chapters discuss ways that consciousness raises problems for our general view of nature, not just for our view of the mind or of traditional cognitive science or neuroscience.

For example, the links between conscious experience, voluntary action, and functional awareness lead to very interesting puzzles when considering multiple personality cases (Braude 1991) or commissurotomy patients (Marks 1981) or blindsight patients (Weiskrantz, 1986, 1988). These puzzle cases can be very seductive, philosophically, but if Liberal Naturalism is correct they are likely more intriguing than they are fundamental. Were we to focus exclusively on overtly cognitive features of consciousness such as these, we would run the danger of confusing the inessential with the essential and of overlooking promising paths in our search.

The history of discovery should lead us to expect the deepest insights to come from reflection on the places of paradox, so, ideally, the features we focus on will yield a paradoxical view of the world when combined with its physical image. The task of removing the paradox-driven tension can provide constraints for our search. Each acts as an explanatory target for a Liberal Naturalist view of nature. They might also provide further clues about the location of our missing knowledge.

In this chapter I make the case that there is a puzzle about how consciousness can exist at the middle level of nature, where it does. In subsequent chapters I discuss whether Liberal Naturalism can plausibly restrain itself to a conservative view that only cognitive entities have experiences. Finally, I examine a set of paradoxes involving such things as the unity of consciousness and the causal relevance of consciousness. I ultimately treat each investigation as producing not just a puzzle or a paradox but also potential clues and explanatory targets. Part II of the volume takes up the task of making sense of these clues and accepts the challenge of meeting the explanatory targets.

4.2 Overview of the Boundary Problem

Bertrand Russell once said that the aim of philosophy is to start with something so obvious as to not be worth mentioning and to end up with something so absurd that no one will believe it. My development of the boundary problem for experiencing subjects is in the spirit of Russell. As with many issues surrounding conscious experience, it takes a bit of hard work to bring the depth of the problem into focus.

I start with the observation that consciousness has inherent boundaries. Only some experiences are part of my consciousness; most experiences in the world are not. Arguably, these boundaries are what individuate me as an experiencing subject in the world. I argue that this poses a problem that any theory of consciousness must answer. How can consciousness have boundaries? What element of the natural world dictates the way these boundaries are drawn? This is the boundary problem for experiencing subjects: We must find something in nature to ground the natural possibility of an experiencing subject bounded in just the way human consciousness is bounded.

4.3 The Foundations of the Problem: Obvious Observations

There are obvious observations that help define for us what it is to be an experiencing subject. First, reflect on the fact that experiencing subjects come in discrete tokens. Without too much strain, we can think of each subject of experience as being a kind of quantum. I am one such quantum, so is Trey Kirven, and so are you.[1] These quanta, the individuated phenomenal fields of experiencing subjects, contain coevolving elements. In some vague but compelling sense of *unified*, these coevolving elements are naturally unified into a subject of experience.

Second, the phenomenal field has boundaries. Not every feeling is part of my phenomenal field because I do not feel the pains produced by damage to your body. The unity and boundedness of the phenomenal field stand together at the core of the concept of an experiencing subject. The driving intuition is that experiencing subjects are inherent individuals in a sense of *inherent* that we must try to make clear. If these boundaries could not exist, then nothing like human consciousness would be possible.

Third, our human consciousness is only a species of experiencing subject. An experiencing subject is a manifold of qualitative entities teeming with variety. We only roughly name these entities in our own case, with such words as *feeling, sensation,* and *appearance.* Other kinds of experience may exist in other kinds of beings.

Fourth, the human subject belongs to a human body and its cognitive processing. Humans, and the activity of human cognitive systems, are individuated at a middle level of the physical world. Typically, our individuation of objects at this middle level of nature is fluid, context sensitive, and interest relative. It is highly conceptual and hinges on facts about the abstract organization and causal cohesion of physical activity. As a consequence, events or objects may form parts of

many individuals simultaneously, depending on how one organizes the world and draws the individuating boundaries. For example, a cell may be an individual; also, at the same time, it may be part of an organ; at the same time, it may be part of an individuated bodily system such as the reproductive system; at the same time, it may be part of the organism as a whole and part of that organism's society; at the same time, it might be part of an ecosystem. For each of these different individuals, a different kind of causal organization exists in the world.

Finally, levels of abstract organization and causal cohesion exist between microphysics and human cognition and between human cognition and the universe as a whole. Lycan (1990) especially emphasizes the importance and continuity of the levels of nature, whereas Scott (1995) has emphasized important differences between them.

All those things should be more or less obvious. The part that is not obvious requires putting these observations together in a way that makes it clear that the experiencing subjects could have been different individuals than they are and different in ways that would prohibit human consciousness. Therefore, we need to explain why experiencing subjects are associated with the specific, middle-level patterns of interaction and organization with which they happen, in fact, to be associated.

4.4 Defining the Problem

The very obviousness of our own existence as middle-level experiencing subjects is an obstacle to appreciating the boundary problem. Because the boundaries of consciousness are something that is always with us, it may not be easy to realize how remarkable it is that things are this way. I now want to bring out the stark brutishness of the fact that experiencing subjects like us could even exist, individuals localized at a middle level, with middle-level boundaries to what we feel. The main points are: (1) if it were not possible to draw these boundaries to the phenomenal field, humanlike experiencing subjects could not exist; and (2) the fact that such boundaries exist where they do is surprising, and their basis is not obvious. To bring out the problem more vividly, I use several thought experiments designed to loosen our sense that there is a natural inevitability to the boundaries that actually exist.

Abnormal forms of consciousness, such as multiple personality disorder (MPD), open the door to the possibility that, in some circumstances, multiple experiencing subjects may coexist within a single brain. Braude (1991) describes cases of MPD in which different personalities may be copresent, each claiming to be a distinct center of awareness. Among many peculiarities, these centers of awareness (which Braude describes as apperceptive centers) make claims to sharing a variety of relations among their experiences. Sometimes they claim distinct experiences altogether. In these cases, the experiences of each personality are "screened off" from the others, so the different personalities achieve, apparently, privacy of experience. In other situations, their experiences partially overlap,

some belonging to multiple centers of awareness and others only to one. In still other cases, experience may be completely shared, although particular experiences may sometimes claim to be owned only by one or another center of awareness. In such cases, John and Mary may both claim to have an experience, but only John claims it as his experience.

These cases raise very puzzling issues about the facts of the matter. What are the number and boundaries of the experiencing subjects that exist in these cases? Is there really one for each personality, or are the claims issuing from confabulation? Whatever the truth, it seems to me that different hypotheses are at least coherent: there could be one experiencing subject, there could be many, and perhaps there are even overlapping subjects of experience. Accepting that there is a legitimate scientific question here, that nature could deliver any one of several possible answers, is a first step in beginning to see the boundary problem.

Most likely, the boundaries of consciousness correspond to the boundaries of certain distinctive activity in our brains. Some evidence suggests that specially synchronized activity in and around the cortex, modulated chiefly by the thalamus, constitutes the boundary maker for human experiencing subjects. For example, Crick (1994; Crick & Koch 1995) hypothesizes that different regions of the thalamus coordinate each level of visual processing. Llinas (1994, 1996) reports the existence of a wave of coherent oscillatory activity that sweeps the cortex every 12-13 msec that is perhaps generated by thalamic activity and postulates that it binds separate sensory content together into a unified representation. Newman (1995, 1997) and Newman et al. (1997) argue that this activity constitutes the binding of sensory contents into a global workspace whose contents are neurally broadcast to specialist subsystems. According to the picture that is emerging, this activity as a whole corresponds to an experiencing subject.

But this raises the questions, What counts as a "whole" for nature, as far as it is concerned with experience? And why? I begin clarifying the importance of these questions by asking, Might any of the subsystems oscillating within this magnificent whole also constitute an experiencing subject? Consider the patterns of synchronized activity that carry and organize auditory information from our ears. Is there an experiencing subject, existing at a different level of nature, that is associated with this activity alone? The picture I am proposing is something like this: Within ourselves as fully human experiencing subjects, there would be other experiencing subjects, themselves perfectly complete subjects of experience, although simpler. Like Russian dolls, there would be individuals within individuals within individuals, all of them subjects of phenomenal experience. The hierarchy of nature might then contain a hierarchy of experiencing subjects, each more or less complex.

The very definiteness of what it means to be an experiencing subject seems to require an answer. Given that we understood the obvious observations about experiencing subjects, extension by analogy should allow us to make sense of this question. Bring to mind the physical image of the world: We have a multitude of

interacting microphysical entities at places and times, this multitude congealing into a macroscopic whirlwind of finely layered patterns of organization. Simply imagine looking at the patterns of physical activity in the world from the perspective of a third-person observer. Note the coherence of causal and abstract organization at the many levels and the many ways it exists. We know that a set of these patterns supports boundaries that allow for the existence of us, where we are one kind of experiencing subject.

The question here is, Does nature support other kinds of feeling subjects, other kinds of experiencing beings? Analogous to us as experiencing subjects, might there exist simpler experiencing subjects whose boundaries are given by subsets of the activity that determines our experience as a whole? After all, the relevant subsets of activity are like the more complete set in many ways. They share common biology with the larger set of events; they carry information and are processing it; they process it in a very similar way; and within themselves, they are internally synchronized and coherent. Do any of these subsets of activity support experiencing subjects, also?

We do not need to assume that an experiencing subject corresponding to auditory activity experiences sound. Instead, it might experience qualia uniquely appropriate to its own level of reality and be responsive to its own finer grained causal organization, just as we are. The case of multiple personality disorder (now called dissociative identity disorder) suggests that a single brain might be able to support several experiencing subjects at the same level of organization. The next step is to wonder whether the normal way of things in the brain might be to support several experiencing subjects but at different levels of organization. The one with which we identify might just be the one at the highest level of organization. Is this the way our world is?

Because such brain activity, taken as a whole, corresponds to the existence of experiencing subjects, *us,* the point seems to generalize to a relatively mild claim. Our intuitive concept of what it is to be an experiencing subject allows for the possibility of simpler experiencing subjects, individuals whose manifold of experience consists of much less rich and less cognitive experience. Given this, the physical activity in fact corresponding to the existence of an experiencing subject might also support other, simpler experiencing subjects via the simpler patterns of organization it contains. Of course, it might not.

I am not claiming anything about the plausibility of this view. I am only claiming that it is an epistemically possible view. It is not a question that one can answer through a priori reflection on the nature of experience and the nature of the physical. It is an empirical question that arises only after one is aware of the physical facts and is suggested by them, as after reflecting on the physical situation, both yes and no seem possible. Why couldn't there be experiencing subjects at many levels of processing, some associated with subsets of the cognitive activity corresponding to our own experiences? On the other hand, why would there be? I now take this openness in the concept and, in steps, parlay it into the full-blown boundary problem for experiencing subjects.

4.5 Sailing toward Scylla and Charybdis

The next step toward the boundary problem is built on a variant of Ned Block's well-known fiction of the Chinese nation simulating the functionality of a human brain (1980). To make it a little less fantastic, we can imagine the simulation of some other, simpler kind of organism's brain, maybe a fish. Very likely, a fish is an experiencing subject.

Imagine building a robot fish. Imagine also that we have designed its nervous system to be functionally isomorphic to the nervous system of a naturally occurring fish (assuming that's possible). The processing has been made remote in the usual way, with inputs and outputs to the fish's central nervous system employing relays. These relays send signals to remote stations manned by human beings. The humans monitor the signals as they come in and relay an output to other destinations. Some signals are sent between the remote stations, and some are relayed back to the fish as motor outputs. In this way, we imagine that the relay system is functionally isomorphic to an actual fish's brain.

The question Block raises is whether or not a system like this system would be conscious. Block uses the example in an attempt to show that our concept of phenomenal consciousness is not a concept of a purely functional entity and that it supports the view that consciousness does not conceptually supervene upon functional organization. However, Block's argument fails to show that the system will not contingently support the existence of consciousness. As is often pointed out, it seems surprising that our brains would support consciousness, but we know firsthand that they do.

Because the relay system is functionally like a fish's brain, it is certainly conceivable that this system actually supports an experiencing subject. The system has parts that are phenomenal homunculi, which is strange, but I previously argued for the consistency of the idea of experiencing subjects whose physical organization supported the existence of other experiencing subjects. In fact, both of these seem to be at least epistemic possibilities:

(1) Each homunculus is an experiencing subject, but the whole system is not.
(2) Each homunculus is an experiencing subject, and so is the whole system.

These possibilities are eye-opening because we can redirect the principles that make them plausible back to a local system for the fish. The homunculi system is functionally isomorphic to the fish's cognitive system. Each homunculus maps onto some important part of the organizational structure of a naturally evolved fish. Imagine the mapping being made with one of the homunculi, call her Edna, mapped onto some functional part of the fish, call it the E-system. There is no principled reason to restrict possibility (1) to the robot fish alone. By analogy, (1) would seem to ground the possibility that the natural fish's E-system, the part corresponding to Edna, could be an experiencing subject, even though the fish as a whole would not be.

How does the analogy go? By admitting possibility (1), we are admitting the coherence of the idea that the robot system as a whole may not be an experienc-

ing subject. In doing so, we are admitting the coherence of a world in which (a) a system may contain experiencing subjects, (b) that system may be functionally isomorphic to the fish's system, and yet (c) that system is not an experiencing subject. In the previous section I gave reasons why it seemed coherent that ordinary cognitive subsystems could themselves be experiencing subjects. To imagine the E-system as an experiencing subject, but not the fish, we have to combine the two points.

To combine them, conceptually shift the boundaries that make experiencing subjects. Shift one's view of nature so that the phenomenal boundaries stretch through the E-system, encompassing all the activity within it but not overflowing the boundaries of the E-system. The larger individual is abolished. In its place is a collection of simpler experiencing subjects in a system of competitive and cooperative interaction. Of course, the experience of the E-system would be vastly different from the experiences of Edna. Accommodating the difference between Edna and the E-system requires postulating alien experiences for some simpler beings. Our ordinary concept of experience is tolerantly open-ended in this way, so this requirement does not stand in the way of our being able to change the intuitively assigned phenomenal boundaries. Lacking clear criteria for natural boundaries, conceptually we can rearrange the boundaries, forcing the individuality down to the E-system level. By doing it, we rob the natural fish of its status as an experiencing subject.

Once we have seen the essential analogy between Edna and the E-system, we can begin to engage in other conceptual shifts. Obviously, the E-system might not be an experiencing subject. It is perfectly coherent to suppose that the only experiencing subject associated with the fish's brain is the one existing at the global fish level. The coherence of the idea that natural fish do not have phenomenal E-systems seems to support a third possibility:

(3) The homunculi system would be an experiencing subject, but none of the homunculi would be.

The possibility that the E-system is not an experiencing subject means that some systems that have experiences in some (epistemically) possible worlds do not have experiences in others. We can apply this principle to Edna. Although everyone knows that in fact Edna would be an experiencing subject, we cannot overlook the failure of consciousness to conceptually supervene on the physical. This failure raises the logical possibility of phenomenal Zombies. Phenomenal Zombies are physical systems organizationally just like human beings but without consciousness. With the specter of Zombies looming, we need to explain why (3) is not true even of our world. For instance, some people think that experiencing subjects emerge at a certain level of complexity, and this seems like an empirical possibility. If it is possible, then imagining (3) merely requires imagining such a world, dictating an appropriate kind of complexity, and then moving the starting point upward, past Edna. The complexity point at which experiencing subjects emerge would be higher than that possessed by the homunculi, but not

by the homunculi fish. The result is that Edna could be a zombie and a component in a system supporting an experiencing subject. What keeps conscious experience right there, between Edna's ears?

This reconception of boundaries is just the flip side of the earlier suggestion. Earlier the reconception was a movement of the phenomenal boundaries to lower levels of organization, robbing wholes of their experiences. Here, the reconception is to a higher level of organization, robbing parts of their experiences. Such a world would be one in which eddies of coherent causation that are human bodies would not support experiences, and human phenomenal consciousness would not exist. We might say that in this imagined world there are humanlike *bodies* but no human *beings*. Instead, the experiencing subjects exist at a higher level. As human bodies act, exchanging signals with one another, as well as interacting with other causal eddies, the phenomenal individual arises only for the supersystem. The possibility is analogous to the way we (or many of us) normally imagine that our cognitive subsystems contribute to our conscious lives without themselves being experiencing subjects.

I intend these science fiction tales to make vivid how an intuitive understanding of experiencing subjects supports a great deal of possible variation in their boundaries. Once the basic point has been appreciated, we can make the point without using philosophers' thought experiments. Even actual systems, such as economies or political systems or nation states, bring it out. An economy is an extremely complex dynamical system of self-organizing components. As a physical system, it stores information and seems to have a kind of distributed memory, a high degree of synchrony between its parts, massive parallel communication, global broadcasting and dominance of certain information, feedback loops, and so forth. The spatial and temporal scales at which this all takes place are much larger than in an individual brain, and much different in detail, but the same basic kinds of activity exist. It would be nothing new to suggest that an economy might represent some kind of group mind. And, really, it is not just a philosopher's question. It is also a deep scientific question about the nature of mind, as well as a legitimate question of fact about something that actually exists in the real world.

What is the main reason for rejecting the idea? The economy certainly has representational properties, and it is a representation consumer: Money, its lifeblood, is a representational vehicle through and through. Mostly, the problem is the bizarreness of believing that the U.S. economy is conscious, and most people consider consciousness essential to mind. Now, I do not know if the economy in fact supports the existence of an experiencing subject and actually tend to doubt it myself. Still, it seems as a priori coherent to me that an economy could (on a much slower time scale) support such existence as it does that my brain would, and I am quite sure my brain does. The economy would have to possess a very different kind of phenomenology, but there does not seem to be good reason for thinking that human-type experiencing subjects are the only kinds that could exist. There's no escaping that economies share many of our mind's most salient characteristics, stretched out vastly in scale over space and time.

Even the scale differences do not amount to much once one considers that our experiences arise from collections of atoms and molecules. The time scale they operate on is far faster than the time scale on which brains produce consciousness. If bunches of neurons (or molecules) stand to us as we do to the economy, and if their organization supports us as experiencing subjects, why couldn't we support the economy similarly? Once we see the possibility that both the economy and our bodies might support experiencing subjects, we are only a short step away from seeing another possibility. It might have been that we are not phenomenally conscious but that the economy nevertheless would be. After all, it seems coherent that our neurons are not experiencing subjects, even though we are. The boundaries of experience, once loosened, can begin to shift radically. Again, why are our bodies not simply local, nonphenomenal causal eddies within a larger phenomenal individual? What grounds the brutishness of these boundaries?

4.6 Scylla and Charybdis: The Boundary Problem

The boundary shifting that occurs in these thought experiments is enabled by the fact that information about physical pattern and organization alone does not fix the boundaries of experience. We individuate most objects at higher levels of organization by extracting some significant pattern from the flux of microphysical interaction. Consistent with a given pattern of microphysical causation, innumerable ways exist of conceiving and reconceiving the abstract organizations that supervene.

Just adding these facts about pattern, or abstract organization, to the causation between the microphysical entities does not seem to go far enough in determining the proper sense of *inherent* in the idea that an experiencing subject enjoys a kind of inherent individuality. One can coherently hypothesize almost as many ways of determining boundaries for experiencing subjects as there are of abstractly organizing and reorganizing the patterns of microphysical interaction in the world. The resulting scenarios are intuitively bizarre, but bizarreness is not inconsistency. The fact that nature's boundaries yield human consciousness stands out as a brute fact.

We are faced with the need to understand more deeply what it is to be an inherent individual in the natural world. We need a natural criterion for individuation, one that illuminates the specialness of some patterns over others as supporters of experience. The fields of the most primitive particles (or strings or whatnot) make one good set of candidates. Each of these has a natural dynamic unity, one that seems inherent. An experiencing subject might be associated with each of these.

This suggestion threatens human consciousness. If the fields of the primitive individuals of physics are the only natural individuals, the rest of us are mere abstractions off the pattern of their interaction. Each primitive physical individual may be a simple experiencing subject, supporting firefly flickers of feeling briefly buzzing at the lowest levels of spacetime, but above them the world is dark. This world would be the panpsychist's world painted surrealistically. There is nothing

that can bootstrap us to human consciousness: feeling, feeling everywhere, but not a drop can think.

Perhaps, by flowing along the lines of interaction, the experiencing subjects could outrun the boundaries of the primitive individuals of physics. Here the trap concerns stopping the flow of interaction. It can seem that the flow of interaction in the universe is inherently unbounded, and no merely abstract pattern presents a natural condition for containing it. Those patterns merely direct it from one watershed to another, orchestrating it, moving it along through the continuity of the universe. According to this view, experience must follow the boundaries to their limits along these lines of interaction. This makes for the possibility of a universal subject of experience, perhaps some kind of cosmic consciousness. Unfortunately, no room exists for the more mundane, middle-level boundaries necessary for human consciousness to exist. Like the first view, this view banishes middle-level individuals from existence.

These two views are a Scylla and Charybdis for Liberal Naturalist theories of consciousness.[2] One view pushes us inward, past the point of middle-level individuation, and into the realm of the subatomic. There, and only there, do we find our natural, inherent individuals. Another pushes us outward, past the boundaries of the subatomic individuals, ever outward along the lines of interaction between them, racing past the middle level to the continuous unfolding of the cosmos. Only there, at the level of the universe, do we find our inherent individual. Neither view allows for conscious human beings. To navigate the middle ground, we must find a principle that allows us to push those boundaries outward from the microphysical but only *just so*. We must be able to go only so far past the microphysical level and no farther. That is the boundary problem for experiencing subjects.

4.7 The Teeth of the Problem: Two Examples

By considering two examples of dual-aspect theories that falter on this problem, we can get a better sense of its importance. The two proposals I briefly critique are the materialism of Michael Lockwood (1989, 1993) and the information theory of David Chalmers (1996). I do not believe that either successfully navigates the way between Scylla and Charybdis.

Lockwood. Michael Lockwood's materialism is a resurrection of Bertrand Russell's neutral monism in the context of quantum mechanics. In an argument similar to my argument from *Life* in chapter 2, Lockwood suggests that phenomenal consciousness fails to logically supervene on the physical because physical concepts are content-neutral, merely specifying the structure of the causal flux. Phenomenal qualities and consciousness, on the other hand, are defined precisely by their content. Lockwood suggests that a nice solution to the problem is to simply draft phenomenal properties into duty as the content of the causal flux whose structure is described by physics. The result is a kind of dual-aspect theory. Physical concepts are about the structural aspects of the causal flux, and our phenomenal concepts are about the intrinsic content that is in flux.

This is an interesting proposal, but something needs to be added before it can hope to account for the individuation of human consciousness at the midlevel. After all, if we are to believe physics, the individuals who are the natural candidates for this basic phenomenal content are the fundamental fields. Lockwood's theory needs to take us from this simple phenomenal content of simple individuals to the complex, middle-level experiencing subject necessary for human consciousness. According to one horn of the dilemma, he is stuck at the microphysical level.

Lockwood (1989) appeals to the other horn for help, postulating that phenomenology flows along the lines of interaction in the world. Unfortunately, he has no principle to allow him to resist being hung on this horn, as it urges that the boundaries be pushed further and further outward. Once the bootstrapping process has begun, Lockwood's theory gives us no explicit way to stop it. Actually, the problem is a little worse for Lockwood, because he is sympathetic to the Everett interpretation of quantum mechanics. Interaction, although structured, is seamless in Schroedinger's world, and Charybdis demands a reason for stopping it here, where there are human cognitive systems in one eigenstate. There does not seem to be a compelling reason to think Lockwood's proposal would result in anything less than a many-worlds-sized individual. Lockwood (1993) discusses this problem and makes this appealing observation:

> In quantum mechanics there is a sense in which all observables, and in particular observables corresponding to every level of structure, are to be regarded as equal in the sight of God, as are different frames of reference, relativistically conceived. As I intimated earlier, quantum mechanics seems to be telling us that it is a classical prejudice to suppose that the world is not intrinsically structured at anything but the level of elementary particles, and their actions and interactions.

This sets out the problem and a possibility for solution. Alas, Lockwood concludes:

> For our own awareness, so I have been urging, embodies a preferred set of observables, which in turn amounts to saying that its contents, at any given time, embody the answers to a set of questions about the state (the intrinsic state) of the underlying brain system. Sadly, however, we here find ourselves in a predicament. . . . We know the answers to those questions, in a way that a scientist, merely by examining our brains from without, never could. But unfortunately, we have, as yet, no idea what the questions are!

In other words, for Lockwood's view to work, we need to find a basis for the existence of an intrinsically preferred set of quantum mechanical observables at precisely the level at which awareness emerges. This, however, is just the boundary problem rearing its head.

Chalmers. David Chalmers (1996) proposes that phenomenal properties and physical properties might be two aspects of information spaces. If we take his suggestion as being unrestricted, it is immediately confronted with the problem of individuating information spaces. On Shannon's view (1948), which Chalmers appeals to, information is a difference that makes a difference along some causal pathway. But a difference that makes a difference to *what* along the pathway?

Falling on the first horn, we can recognize informational differences to the basic individuals, but that banishes human consciousness.

Falling on the second horn, we can recognize the universe as a whole as an information space. Its structured state changes as interactions occur within it, and one state of the universe makes a difference to subsequent states, but this also banishes human consciousness. On Chalmers's proposal we should be able to save middle-level individuals by allowing for all covarying subportions of spacetime to be information spaces, but then we are left with panpsychism run wild. Even within one brain, we will have astronomically large numbers of experiencing subjects, separately experiencing, each corresponding to different ways of carving up the activity of the brain and its causal pathways. An explanation so promiscuous is not illuminating.

4.9 What to do?

The Liberal Naturalist should take the boundary problem seriously and think hard about what might be missing from our current view of individuation in the world. The suggestion that we allow inherent individuality to flow along the paths of interaction between individuals sounds promising. After all, we are looking for something more than abstract organization to ground judgments of natural individuality, and causation seems to be an inherent, natural connection par excellence. Also, an interaction divides the world by its very nature, partitioning it into different spaces that mutually condition one another. Lockwood's observation suggests that the second horn of the dilemma gains its conclusion by taking advantage of a naive view of interaction, one that capitalizes on a rough classical understanding of causation.

One good strategy to follow would be to think harder and more carefully about interactions in the world. We should think in more detail about the way they might condition nature into individual, mutually influencing regions, and do so at many levels simultaneously. We very well might discover that interactions have certain important aspects we can use to mark off candidates for natural individuation. These individuals would then be candidates for supporting experiencing subjects. The job would then be to look for a physical reflection of this special feature of interactions.

Like Lockwood, I think that we must understand the causal structure of our world better. Causal connections seem the best candidates for helping to understand more deeply the naturally individuated, middle-level structure exhibited in our phenomenal existence. This is the first example of a conclusion that pops up again and again in this section of the book. By its end, it will seem that wherever we turn in trying to understand consciousness, we end up spun around and facing questions about causation. This section of the book thus serves not only as a discussion of problems and challenges facing Liberal Naturalism but also as a runway to the eventual topic of causation.

5

On the Possibility of Panexperientialism

5.1 Introduction

According to Liberal Naturalism, consciousness shows us that our world has another *fundamental* aspect that we must understand if we are to understand the qualitative character of our mental lives. To make sense of this fundamental aspect, Liberal Naturalists have to introduce some extraphysical fundamental natural laws or principles. These laws or principles will either govern the behavior of experience directly or govern the behavior of something nonphysical that underlies experience. Because the laws or principles governing this aspect are fundamental, it is not pretheoretically plausible that neuroscience and psychology are the right places to look, much less the only places to look, for clues about them. We have to consider looking beyond issues in psychology and the philosophy of mind for clues.

When considering potential new fundamental laws responsible for the existence of experience, there is always the danger that we might find reasons for thinking that instances of experience actually outrun instances of cognition. Although such a thing would not be acceptable as an ad hoc hypothesis brought in especially to explain experience, it might be a natural consequence of an independently motivated view. The view that experience outruns cognition is called *panexperientialism,* a term introduced by David Ray Griffin (1997). Panexperientialism is the view that experience exists throughout nature and that mentality (i.e., a thing requiring cognition, functionally construed) is not essential to it. It is a milder form of traditional *panpsychism*, which is roughly the view that everything has an experiencing mind associated with it. Is panexperientialism even possible? A Liberal Naturalist cannot lightly dismiss panexperientialism and so needs to reflect on its strengths and weaknesses, weighing them appropriately. In summary:

1. Liberal Naturalism needs to posit extraphysical fundamental laws or principles to explain the relation between the physical facts and the facts of experience.
2. The simplest and most fruitful theory of those laws might have panexperientialism as a consequence.
3. Therefore, the Liberal Naturalist needs to understand whether panexperientialism is even a theoretical option.

In this chapter I undertake an extended, critical reflection on the viability of panexperientialism and whether it is even a possibility that our world might be a panexperientialist world. In this discussion, I assume that every subject of experience possesses a field of experience containing a variety of phenomenal qualities. I call this phenomenon the *qualitative field. Qualitative* is just meant to capture its close relation to qualia; perhaps it subserves them in some way. I use *field* to denote a bounded collective. Our phenomenal lives contain many distinct qualia: itches, sounds, smells, emotional tones, and tickles are examples.

Here is the picture. Nature merges all these different qualia into one subject of experience, individuated from other qualia not only by their type but also by the *field* of experience to which they belong (e.g., mine and not yours). The boundaries of this field individuate *subjects* of experience by including and excluding feeling. We can think of the unified, bounded collection of qualia that constitutes the experience of an individual as the qualitative field associated with that individual. Liberal Naturalism must face the problem of providing a basis in nature for the existence of such a thing.

From this point forward, I use the term *cognition* to refer to functionality of the brain, including basically everything studied within cognitive science. Where I use other psychological terms, such as *memory*, *conceptualization*, or *perception*, in association with *cognition*, I mean the purely operational sense these terms have within cognitive science.

I am not trying to be contentious by using *cognition* this way, especially to people who are convinced antiphysicalists. I am only making a concession to clarity, not an endorsement of any theoretical view. By making this concession, I can more easily show why we might need to go beyond issues specific to psychology and the philosophy of the mind in formulating a theory of consciousness.

5.2 Is There Evidence for Panexperientialism?

The idea of panexperientialism, much less its truth, sits poorly with some people, who usually suspect that it is incoherent. There *are* two serious intuitive reasons for rejecting panexperientialism outright: (1) we have no evidence for the existence of experience outside of cognitive contexts; and (2) the mere supposition is incoherent because divorcing experience from cognition requires experiencings without appropriate experiencers. The question here is whether these intuitive reasons withstand scrutiny.

We can reject reason (1) simply by noting that every theory about conscious-

ness goes beyond the direct evidence that we have, because we have direct evidence only in our own cases. From my own perspective, any theory that attributes consciousness to people other than myself is going beyond my evidence for the existence of consciousness. More generally, what I count as evidence for attributing consciousness beyond my own case will *depend* on my theory of consciousness. Therefore, the concept of *going beyond the evidence* is poorly defined.

For example, if I believe consciousness depends on language ability, then verbal reports of conscious states are the only kind of "evidence" that my belief will allow me to recognize. Under the influence of such a theory, I would deny consciousness to animals, because we have evidence for consciousness, via verbal reports, only in people.

Alternatively, if I believed that consciousness depended primarily on biology, then I would extend attributions of consciousness down the phylogenetic chain to other animals. The basis of the extension will be a claim that the common biology we share with them is evidence that they also are conscious.

Similarly, if I adopted the stance that only cognitive functioning such as the global availability of information is relevant to the presence of consciousness, then I would, based on the evidence, suppose competently functioning silicon robots to be conscious. The moral is that what we count as evidence for consciousness and our theory of consciousness are heavily intertwined. Thus no pretheoretical bias about the evidential base can carry an overriding veto power on the form of the final theory.

Surely, the theory must *include* certain systems as conscious (any theory that had the result that only Gregg Rosenberg was conscious should give us pause). In general, our theory should include people, and expecting it to include other mammals, fish, and birds is reasonable, also. Things get a bit fuzzy when we begin to consider insects, perhaps, and also artificial systems, but that is acceptable.

Initially, our evidence is strongest about certain kinds of organisms our theory must *include*, and it is much weaker about any kinds of systems a successful theory must *exclude*. We should, therefore, concentrate primarily on finding the simplest, best motivated, most coherent set of laws that include the systems that intuitively should be included. If an otherwise exceptional theory has the consequence of also associating experience with some surprising class of systems, then we should accept that consequence as a discovery about nature as far as consistency with the evidence is a concern.

5.3 Is Panexperientialism Coherent?

The second reason for insisting that cognition is essential to experience is the fear that any alternative is incoherent. This is probably the main reason for rejecting panexperientialism out of hand, and it is the more difficult hurdle to clear. Although I am not sure one can ever *fully* shake this intuition, there are ways to lose confidence in it. The consideration most able to undermine the intuition is

that only certain kinds of qualia force themselves on us as essentially mental, and other kinds, with very different characters from those we know, might subsist outside of minds.

In detail, my reply to the second intuitive objection breaks into four observations:

1. Our concept of experience is highly open-ended and can be sharpened to either include or exclude noncognitive experiencers.
2. The experiences we might attribute to noncognitive systems do not contain "little pains" or "little specks of blue" but instead have some kind of qualitative character very alien to us.
3. The best way to conceive of those qualitative fields is via a mental place holder for the solution to the analogy problem, "Y is to system X as experience is to the human mind," which sets up Y as a qualitative experience that we know might exist, but which we cannot concretely imagine.
4. The best term for the alien character of these fields is *protoconscious*, a term meant to suggest that they contain experienced qualitative objects that are not, strictly speaking, being experienced by a mind (because there is no associated cognition). These protoconscious states are states of pure experience. They need not have semantic content, and certainly no cognition will occur within the manifold of experience.

The open-ended character of experience. Observation 1 appeals to the open-ended character of our concept of experience. The privacy of consciousness forces us to build in a kind of tolerance for alien experiences and feelings: A manta ray sensing the electromagnetic structures on the ocean floor may experience qualities we could never imagine. We also have to allow that simpler and simpler organisms may have experiences of simpler and simpler kinds, as well as alien kinds. So the open-ended character of the concept requires us to accept that there could be experiences both very alien to, and much simpler than, any we can imagine.

On reflection, it is uncertain how far from our native experiencings the concept of experience could be extended and remain viable. It seems to taper off vaguely and may carry very far from its point of origin. Although this open-ended character may put us on the slippery slope to incoherence, the slope could just as easily turn out to be harmless. For all we know, noncognitive experiencing subjects, although odd relative to our kind of experiencing, are perfectly coherent kinds of entities. At the very least, the open-ended character of the concept means that we do not actually have a definitive reason to reject the idea, and, depending on our needs, we can sharpen the concept of experience in a way that includes such entities or in a way that does not. From our vantage point as theory makers, it seems to be up to us.

The possible existence of alien experiencings. Observation 2 is that the experiencings of a simple noncognitive system do not need to be of simple, structureless elements found within our own consciousness. Someone might believe that

it *could* be that way, but we do not have good reason to believe it *has* to be that way. To conceive of the alternative, imagine that all phenomenal individuals experience in a way that corresponds to the causal contributions that their physical components make to their physical state. For example, consider a person and an event in that person's cognitive subsystem contributing an element to experience. The evidence we now have suggests that the character of its contribution will correspond in some way to the informational difference the cognitive subsystem makes to the person's overall state of being.

As I discussed in chapter 4, that cognitive subsystem might be a subject of complex experience itself. The contribution that components within its own rich, internal causal structure make to its overall state would likewise determine the character of its unique experiencings. The panexperientialist possibility, if it is a possibility, is that some truth like this holds deeper and more widely than most of us would have guessed: that for many kinds of systems, not just cognitive systems, there are subjects of experience whose experience is determined by their causal organization.

Why couldn't our world be this way? When we speak of the qualitative field of some other, noncognitive, system, we are obviously not attributing to it the qualities of our own experiences. We are not attributing little pangs of pain or experiences of tiny blue dots to noncognitive systems. Whatever we are attributing, it is not any kind of feeling with which we can empathize. We are supposing that there are experienced qualities that share some essence with the qualities of our experience but that are not cognized and perhaps do not support certain properties useful only for cognitive purposes (such as intentional properties).

The essential analogy. Instead of generalizing our own phenomenal properties onto nonmental entities, the panexperientialist is attributing to these entities an experience that has a character in some very abstract sense *like* that of our experiences but *specifically* unimaginable to us and unlike our own qualia, which brings us to observation 3. In trying to gain an initial understanding of the panexperientialist claim, our best tactic is to maintain that the properties in question are a placeholder for the solution to an analogy problem; for example, "X is to a film plate as conscious experience is to the human mind,"[1] where we know X must have a solution in nature but we do not really know what that solution is. It is an existential claim whose instantiation is something that we cannot be acquainted with and hence should not pretend to understand fully.

Protoconscious properties. The preceding considerations begin to undermine my confidence that the panexperientialist hypothesis is incoherent. With concepts as open as experience and feeling, I cannot decide a priori that the world is not a panexperiential world. If we can assign some sense to the proposition that the cognitive producer (i.e., the mental subsystem) of a feeling of pain, or a tickle, could exist as an experiencer in itself, an experiencer that contributes to human awareness but is not dependent upon it, then I need some empirical reasons for ruling it out. By extension, it seems I need empirical reasons for ruling out the

panexperientialist hypothesis. Collectively, what observations 1-3 suggest is that the difficulty of imagining noncognitive experiencings comes from a kind of cognitive rigidity and not from a fundamental conceptual incoherence. More tellingly, these observations reinforce some further considerations and, together with them, provide much more confidence about the underlying coherence of panexperientialism.

For example, even if some sort of panexperientialism is true, we should not naively assume that every perceptual or conceptual individual, such as a thermostat or a rock or a film plate, has experiences. Large-scale, enduring, coherent experiencers may be extremely rare. As a dilution of traditional panpsychism, the panexperientialism we end up with may be as benign as would occur if the interactions between very simple atoms or molecules mainly produced flashes of extraordinarily simple and brief feeling, like fireflies quietly flickering in the night. For these reasons, referring to the experiences of noncognitive systems as protoconscious rather than conscious is really best.

Even without a cognitive engine being present, there may be a perfectly good sense in which each feeling or protofeeling is part of a subject of experience. By saying this, I am just pointing out the panexperientialist suggestion that not all subjects of experience are cognitive (and hence *mental*) systems.

These conclusions also suggest that both parts of the etymology of *panpsychism* are misleading. The *pan* in *panpsychism* is misleading because it will not be the case that *everything* has experience. By assumption, only some feature(s) of the world correspond to experiencing systems. Even if some noncognitive systems may have experience, the theory will still *constrain* which kinds of things will be experiencers, and the ultimate position may end up being milder than is often feared. As I remarked previously, rocks need not experience, nor thermostats, even if some variety of panexperientialism is true. (I follow Griffin in retaining the *pan* only because I feel *neopanexperientialism* is an uglier name than any position deserves, but I want to retain something suggesting the radical nature of the hypothesis.)

The *psychism* is misleading because one need not associate experiencings exclusively with cognitive activity and hence not exclusively with *minds*. Therefore, even if panexperientialism is true, it does not follow, without further assumptions, that mentality outruns cognitive-style functioning.

5.4 Protoconsciousness and Representationalism

I intend the term *protoconscious* to suggest the hypothesized kinship between the quality of experience for noncognitive systems and our own experiences and also the alienation from its richness, variety, semantic significance, and cognitive awareness. The properties of protoconsciousness can be usefully and explicitly contrasted with the *protophenomenal* properties proposed in Chalmers (1996). According to Chalmers, protophenomenal properties would be fundamental nonexperiential, nonphenomenal properties. By hypothesis, in proper combination pro-

tophenomenal properties could become experienced phenomenal properties. Chalmers leaves open what contexts can provide the proper combination, but we can presume only cognitive contexts work because the proposal seems designed to avoid panexperientialism. In contrast with protophenomenal properties, the properties of protoconsciousness are experiential properties properly considered phenomenal, but they do not require an associated cognitive engine to be experienced.

Because properties of protoconsciousness can be experienced by entities without cognitive engines, it is natural to suppose that they might not have certain features, most especially representational features, that people sometimes argue are essential to phenomenal properties. I call any view holding that representational features are essential to phenomenal properties *representationalism*. If it is true that properties of protoconsciousness would not have representational features, and if representationalism is right, this would preclude properties of protoconsciousness from being phenomenal properties. It would then be more arguable whether protoconscious properties could be experiential properties at all, and the panexperientialist's view would be more doubtful than so far supposed.

To be truly viable, the panexperientialist position must have an answer to this representationalist challenge, and I believe the panexperientialist has several viable replies. I present these replies in what I believe is the reverse order of attractiveness, from the least attractive option to the most. I emphasize that *attractiveness* here is my subjective judgment of their plausibility, fruitfulness, and simplicity relative to one another, and not a judgment of their in-principle adequacy (I think they are all adequate) as responses to the representationalist challenge.

Reply 1: Inert intentional features. One answer to the representationalist challenge could be that the properties of protoconsciousness *do* have intrinsic intentional features but that these features are inert unless the associated experiencer is a representation consumer. On the assumption that only cognitive engines are representation consumers, then outside such contexts the intentional features of a protoconscious property would be like the electrical charge of an ion when surrounded by materials with which it cannot interact. On this reply the properties of protoconsciousness are phenomenal properties even by representationalist standards, and there is no dispute.

Reply 2: Representationalism is problematic for the human case. A second answer to the representationalists rejects their fundamental proposal that all the phenomenal properties in human consciousness are representational. The panexperientialist can observe that the representationalist position rests on certain paradigm cases, such as the visual experience of shape, for which it is compelling (for normal subjects) to conclude that *those* phenomenal experiences essentially represent certain spatial features. However, the representationalist generalizes these cases to claim that *all* phenomenal experiences are essentially determined by (or identical to) representational content.

The representationalist case for the generalization is far weaker than the case

for the paradigm examples. It requires (1) a theory of representation that can be applied to all experiences, normal and nonnormal; (2) a consistent and plausible application of that theory of representation to a truly representative variety of experiences that implies that each is determined by (or identical to) some representational content; and (3) a case that each of these experiences is tied *essentially* to its representational character.

The panexperientialist would be right in maintaining that no representationalist has come very close to meeting all three of these criteria. For example, several authors have proposed teleological theories of consciousness (Lycan 1996; Dretske 1995; Tye 1995, 2000) based on reductive theories of representation. On these views, the representationalists theorize that the character of conscious states is completely determined by (and even identical to) the intentional content of a teleological representation.

The evidence that these reductive theories of representation cannot meet condition (3), which states that the tie to representational content has to be *essential* to phenomenal properties, is clear from the fact that the proponents of teleological views always appeal to the unexplained and previously rejected[2] notion of metaphysical necessity to connect the intentional content of a representation to its qualitative character. Without this appeal, there is no plausible argument that the proposed representational content, which comes from a vast extrinsic history of the organism's species, is essential to the connected phenomenal properties.

Others such as Siewart (1998) produce distinct arguments for a representationalist conclusion. These views usually appeal to internalist concepts of representation in which phenomenal experiences nonreductively and intrinsically represent. However, these arguments are based entirely on phenomenological evidence from a limited number of examples in which the phenomenology is compelling for normal subjects. When it comes to condition (2), applying it consistently and plausibly to a representative variety of examples, there are still serious open questions that provide obstacles.

Consider again the example from the beginning of chapter 1 of this book: When we close our eyes and place our hands over our eyelids, many of us will experience diffuse and jumpy patches of diluted color appearing, disappearing, and floating in the darkness. These patches of color are not attached to the surfaces of any perceived objects, do not provide a guide to behavior for the representation consumer, and are not taken as representing any properties by the representation consumer. They can actually be experienced consciously as *not* representing anything, as "pure experiencing" by someone inclined to take them that way. I believe that for a Liberal Naturalist considering representationalism, representationalism about such states can plausibly be denied.

Even some very central and standard phenomenal properties, such as scents, pose problems because not all particular scents seem to correlate in a representational way to specific external properties. In fact, because of the cross-modal way that scents are coded in the brain, there is good reason to think that there are standard significant variations in the scent space between different individuals. If

true, this would make it very difficult to make a case that scents intrinsically represent external properties.

Synesthesia. Representationalism seems to lose plausibility when we look at empirical evidence about nonnormal experiencings such as in synesthesia (e.g., Wager, 1999). Synesthesia is a syndrome in which the normal experiencings of a stimulus include qualia from what are, to normals, different perceptual modes (Cytowic 1989, 1993, 1995). The most common kind of synesthesia is a syndrome in which numbers and letters are experienced as having distinctive colors. For example, a synesthete might always see the number 2 as being green and the number 4 as being red. Synesthesia is real and highly reliable, showing up early in life and being stable across decades of life. It is also hereditary and can be passed from parent to child and shared among siblings. Finally, the evidence for it is clinical and neurological, as well as testimonial.

Letter/color and number/color synesthesias are far from the only kinds. Another kind is sound/shape or sound/color synesthesia, under which the perception of a sound is always accompanied by a color percept or shape percept. Importantly, these percepts are not simply imagistic danglers on the perception but integrally bound up with it. For example one synesthete with the initials DS describes what it is like to have a cross-modal perception of sound that includes shape in the following way (Wager 1999, taken from Cytowic 1989),

> DS: The shapes are not distinct from hearing—they are part of what hearing is . . . That's what the sound is; it couldn't possibly be anything else. (p. 65)

This testimonial should give pause to representationalists. One of their chief arguments, and probably the strongest of their arguments, has been relying on certain central examples where it seems impossible that nature could vary phenomenal content without varying representational content. DS's testimony is in direct contradiction to one of the representationalists' most compelling examples: the supposed inconceivability of divorcing the phenomenal experience of shape from its representation of spatial properties. DS claims that he not only conceives this disassociation but also experiences it. To him it seems impossible that certain shape properties do not form part of the representation of *sound*. If these shapes represent at all, they perhaps represent certain of the aural properties of sound. Perhaps shape experiences are essentially geometric only, and the properties, if any, that our cognitive systems use their intrinsic geometric structure to represent can vary. On some uses they represent the geometric properties of space, and on other uses they represent geometric properties of sound.

Descriptions such as DS's raise danger signs about how our potential judgments of essentialness could be corrupted by cognitive rigidity. Perhaps both our and DS's inability to conceive how certain disassociations between representational contents and experience may occur are due to cognitive rigidity. Cognitive rigidity could limit our conception of how nature is able to use a given phenomenal quality and so delude us into false intuitions of essentialness. If we do not

take this humble approach, we need to justify why accepting the impossibility statements of "normals" rather than synesthetes is more than a simple bias for our own cases.

These cross-modal associations can be useful, as well. Cytowic reports that synesthetes as a group have superior memory, due to the usefulness of cross-modal representation in providing a basis for mnemonics, as well as the fact that synesthesic experiences have stronger emotional content than the experience of normals. They also have greater than average intelligence as a group. There are other uses for cross-modal experiences, as told by Carol in the following account of visiting the dentist. Carol is a synesthete with several syndromes, one of which is experiencing pain as orange (http://web.mit.edu/synesthesia/www/carol.html),

> One example of synesthesia being distinctly unpleasant: I was at the dentist, and he was drilling. And I don't like the sound of the drill – but the color orange that completely flooded my vision, I couldn't shut my eyes, because they were already shut! [laughs]
>
> Except that I'm able to use it diagnostically. I had to have a root canal done once (not my favorite game) but you know, sometimes when you have a tooth pain you're not quite sure which tooth it is? He said, "I can't really say that you need a root canal in this tooth." I said, "This tooth is orange; please do it." And he hesitated. I said, "Look. If I'm wrong, this tooth will never need a root canal." So he went ahead and did it.
>
> He said—he poked around a little bit—"This tooth needs a root canal." He said, "It hasn't really become 'ripe' yet, but the nerve is dying." And sure enough, when the nerve was out, and the anesthesia had worn off, there was no more orange. It's like orange is my default color for pain.

Other synesthesiac experiences covary specifically with properties of the stimulus. For example, MM, a sound/color synesthesiac relates (Wager 1999, from Cytowic 1989),

> The only real problem is that when I am driving and a very loud sound comes on such as loud music or the Alert Test tone and it is hard to see. The image intensity is directly proportional to the sound level. People laugh when I say, "turn that down, I can't see where I'm driving." (Cytowic p, 51)

Finally, Carol uses the number/color aspect of her syndrome to illustrate the heritability of synesthesia with this wonderful story:

> I came back from college on a semester break, and was sitting with my family around the dinner table, and—I don't know why I said it—but I said, "The number five is yellow." There was a pause, and my father said, "No, it's yellow-ochre." And my mother and my brother looked at us like, "this is a new game, would you share the rules with us?"
>
> And I was dumbfounded. So I thought, "Well." At that time in my life I was having trouble deciding whether the number two was green and the number six was blue or the other way around. And I said to my father, "Is the number two green?" and he said, "Yes, definitely. It's green." And then he took a long look at my mother and my brother and became very quiet.

> Thirty years after that, he came to my loft in Manhattan and he said, "you know, the number four is red and the number zero is white. And," he said, "the number nine is green." I said, "Well, I agree with you about the four and the zero, but nine is definitely not green!"

So here we have occurrences of phenomenal properties that are (1) stable across individual lifetimes; (2) heritable across generations; (3) useful; (4) in some instances, variable with properties of their stimulus, even though the stimulus is not "normal"; and (5) in other instances, not variable with properties of the stimulus, but it is inconceivable to the experiencer that they could experience the stimulus without it. I am aware of no remotely plausible theory of representation that has been put forward under which these synesthete experiences turn out to be representational yet do not turn out to have nonstandard representational contents (or at least would be able to have nonstandard contents, given the right story about natural and sexual selection). For example, the view of representation I favor is an action-oriented view of representation in which the representational content of a mental entity is determined by the way it provides guidance to action. Under such a view, synesthesia is a counterexample to representationalism.

Pointedly, a minimal representationalist claim is a supervenience claim:

> $R_{supervenience}$) Representational contents necessarily determine phenomenal contents.

Stories like Carol's visit to the dentist or MM's trouble seeing through a loud noise are very strong indicators that a given representational content ("damage here" or "audio intensity high") can yield different phenomenal contents (no color content in a "normal" but color content in Carol and MM). If true, the evidence falsifies the supervenience thesis. Furthermore, for representationalists to make trouble for a theory on which some phenomenal properties fail to represent, they have to go beyond supervenience and claim essentialism:

> $R_{essentialism}$) Phenomenal contents necessarily determine representational contents.

However, $R_{essentialism}$ must come to grips with testimony such as DS's. DS testifies that a phenomenal property such as shape is perceived as essentially bound up with the perception of a nonstandard object, such as a sound. Yet it is not completely clear that DS's phenomenal experience of shape represents any property of that sound (and if it turns out to represent some aural properties, there are other synesthete examples that could be substituted). The phenomenal quality is not *diaphanous*. That is, the synesthete does not "see through it" to the property of an object but can focus on the quality itself.

In synesthesia, the binding of the extra quale with the object occurs at midperceptual levels of processing. For example, imagine a letter/color synesthete who always sees the letter *d* as green. If this synesthete is shown a block of numbers and letters and is asked to identify how many occurrences of the letter *d* are in the block, she will be able to make the identification far more rapidly than a non-

synesthete would. This is just the response a normal person would have if the *d*'s were colored and the normal was able to use the color as a rapid way of identifying which of the letters were *d*'s. The plausible explanation for the synesthete's ability is that she is using this same recognition strategy and the *d*'s in the chart are "popping out" at her because they are in fact experienced as colored.

In these types of experiments, the association of color with letter is clearly occurring prior to conscious recognition of the letter type but after preconscious object recognition in perceptual processing. It is occurring, by all evidence, at a middle level of processing where perceptual systems are binding qualities to objects once those objects are distinguished by other preconscious systems. This suggests that the processes binding the color to the letter are responding *specifically* to the results of the processing stream in charge of object categorization of the letter and *not at all* to the results of the processing streams responsible for categorizing surface reflectance. Further evidence of this interpretation comes from the fact that letter/color synesthetes commonly can still recognize the true color in which the letter is printed, suggesting that identification of surface reflectance is occurring normally and separately. This looks like a prototypical exaptation story, in which a color-binding mechanism adapted for one function is co-opted to perform another.

If so, the association of the color with the letter is not a misrepresentation of surface reflectance but an accurate response (for the synesthete's perceptual system) to the categorization of the object as the letter *d*. This is a stumbling block for representationalism, as its strongest response to the synesthete cases is to hold that preconscious perception in these cases is misrepresenting surface reflectances. The actual processing story behind production of the experience does not seem to support that interpretation.

Furthermore, I submit that any defense of $R_{essentialism}$ must be nonbiased and noncircular. This means that a legitimate response cannot assume the normal's intuition that an experience of shape by a synesthete like DS inherently represents a spatial property despite DS's testimony. Doing so and then using that to deduce or suppose that there is, after all, an imagistic or illusory representational content of a spatial property within DS's experience of sound does no more than take one of the rigidities in the normal's own cognitive architecture and assert it as a universal truth. All things considered, it seems that synesthesia poses tough challenges for representationalism. I believe these problems are more than serious enough to provide justification for not accepting the representationalist claims.

Reply 3: No basis for generalization. Finally, the answer I most favor to the representationalist challenge is to appeal to observation (2) that began this section of the chapter:

2. The experiences we might attribute to noncognitive systems do not contain "little pains" or "little specks of blue" but instead have some kind of qualitative character very alien to us.

At most, the representationalist can appeal to the phenomenal properties that occur as part of human experience. No one disputes that the human cognitive engine consumes representations, and it is not surprising that contents so closely correlated with its activity are representational. However, even if it were to turn out that *all* phenomenal contents in human experience were representational and *essentially* representational, that would not provide a sound inductive grounding for generalizing about all families of possible phenomenal properties. From the point of view of inductive logic, the problem is obvious: The samples are all being taken from a highly biased set. All the investigated phenomenal properties are, of operational necessity, sampled from those associated with a representation-consuming cognitive engine. It is analogous to collecting all one's information about electric pulses by tapping phone lines and then concluding that electric pulses essentially represent sound and data. Liberal naturalists should be methodological purists, and, faced with an otherwise attractive theory, there would be no methodologically sound way to generalize from this biased sample to proposed limits on other possible families of phenomenal properties.

In particular, one can imagine specific conditions in which good methodology would recommend accepting nonrepresentational phenomenal properties. Imagine a future theory of phenomenal properties that is most simple and fruitful only if it has panexperientialist consequences. In accord with observation 2, imagine that it proposes qualitative characters alien to those associated with the human cognitive engine. In the imagined theory, the phenomenal properties will have some specific job to do—there will be a clear and coherent reason they are associated with the systems with which they are associated—and their essential features will be specified according to the general requirements of their job within the proposed theory. By hypothesis, representation will not be one of those essential features. *If* a theory meets these conditions, methodology demands that Liberal Naturalists should accept the more general and clarified theory over an inductive generalization from a small biased set of highly specialized phenomenal properties.

Summary. In the end, perhaps the kind of high-level, conceptualized experience we find in ourselves is a rare variety of experience. Appealing again to the metaphor I used previously, perhaps most experiencing entities are much closer to the ground level of reality, little fireflies in the night supporting brief flashes of sensation as they interact. This relatively benign state of affairs could be the state of affairs in our world, and I believe it is counterproductive to prejudge the possibilities and, thereby, be tempted to overreact.

6

On the Probability of Panexperientialism

6.1 Why We Must Go beyond the Mind

Are we to think of experience as an artist's flourish? As evidence for how uncomfortable the problem of placing consciousness makes us, observe how even antiphysicalists such as Chalmers (1995, 1996) sometimes pose the problem as understanding how "experience arises from physical processes": that language pushes us to think of the physical as primary, with whatever aspect of nature consciousness belongs to viewed as an add-on. It just "arises."

A more satisfying result would be a *deeper* view of nature that somehow gets under physics. The fact that we need to get under physics is obscured by this intuition about consciousness "arising," which in turn is lent crucial support by the intuition that experience exists only in cognitive systems. Therefore, understanding whether this phrasing is a blinder, and removing it if it is, is an important move in learning to properly appreciate the scope of the challenge.

The antiphysicalists have taken a first step by realizing that the problem of consciousness goes beyond understanding the structure and function of physical systems. This has allowed them to bring the hard problem of experience into sharp focus, highlighted against the background of the problems concerning function and structure. The Liberal Naturalist must be willing additionally to tease apart the problem of experience and feeling from problems of mentality and learn to see the more general problem of finding the basis in nature for qualitative content.

That is, even the hard problem of consciousness may be two problems superposed. It may be a general problem about finding a basis for qualitative fields,[1] their place in nature, and the laws governing them. Also, it may be a specific problem about how the influence of cognition can give a qualitative field the character of *consciousness*. In the last chapter, I argued that *in principle* these are

separate problems. In this chapter, I argue that *in fact* these (probably) are separate problems.

This is a strong conclusion. Most of us possess a natural and strong opposing intuition that these qualitative fields are quite special things peculiar to minds. Common sense suggests that, because systems are only mental when they support cognition, qualitative fields must *arise* within cognitive systems alone. If a Liberal Naturalist wants to press this commonsense assumption about the connection between experience and cognition, things get a little murky because the criteria for a system being cognitive is itself so unclear. It would be cheap to make nearly *everything* cognitive simply by definition. So we can, for the sake of argument, simply stipulate that the class of cognitive systems supporting experience must be like human cognition in certain sophisticated respects.

An example of the sort of high-level account I have in mind comes from Robert Kirk (1994, 1995), whose proposal exemplifies the intuition very well. He proposes a set of information processing capacities that he calls *the basic package*. The basic package includes:

- The possession of needs and goals.
- Self-directed behavior.
- The capacity to acquire information which is *for* the system in the sense that it can pursue goals by modifying its behavior in response to the information.
- The ability to store information.
- The ability to assess information.
- The capacity to make decisions.

Kirk suggests that the elements of the basic package are the necessary and sufficient conditions for the existence of consciousness. This is precisely the kind of view targeted by the following discussion because it represents right-spirited attempts to restrict definitions of *cognition* just enough to save consciousness for a class of intuitively or pretheoretically acceptable systems. From this point on, I refer to cognitive and noncognitive systems to denote the vague classes of intuitively acceptable and unacceptable systems.

In the following I explore the difficulty of finding appropriate fundamental laws that meet a set of constraints like Kirk's and how unlikely those laws will seem. The key question concerns what *class* of base properties would guarantee the existence of consciousness on such a theory. There are three popular possibilities: the level of complexity of the system, the kinds of functionality possessed by the system, or one of those plus biology. I argue that each of these seems implausible as a feature in a fundamental law of nature.

Even now we can imagine the coarse-grained forms such laws must take, and we have some metacriteria for judging potential laws: We can judge the appropriateness of the concepts they employ; we can speculate about the simplicity of the resulting laws; and we can speculate about possible empirical consequences with an eye toward judging the plausibility of such consequences. The purpose of

each discussion here is to see whether our putative fundamental laws would have a character or consequences less implausible than panexperientialism itself. The general form of argument in each following section is:

> The proposed base properties imply problems regarding complexity and implausibility of the corresponding fundamental laws.
>
> Accepting these problems just to avoid panexperientialism would require setting aside standards of good theory construction.

The discussion in this chapter is of immediate practical importance and not merely speculative. In chapter 12 I propose a fundamental law governing the existence of experience. The form I choose for this law is very simple and straightforward, and it proves to be fruitful: From this law we will be able to derive many of the specific features of experience.

However, the fundamental law I introduce has panexperientialist consequences. I must make a methodological choice at the point where I introduce it. I could choose to avoid panexperientialism by breaking the law into a disjunction, where one of the disjuncts would tie experience solely to the "right" kinds of systems, using the other disjunct to propose some unknown properties (perhaps "protophenomenal" properties) that do the required work outside of cognitive contexts. It is there, at that choice point, that the following arguments show their real bite. Considerations such as the following make it clear what an undesirable choice it would be to replace a simple and straightforward law with a law jury-rigged to give the "right" result outside of cognitive contexts.

6.2 Complexity

The first suggestion is that some fundamental law of nature guarantees that a qualitative field becomes associated with a system when it reaches a certain level of complexity.[2] The coarse form of the fundamental law would have to be, *If a system reaches level of complexity N, then a qualitative field must arise from and coevolve with it.* Of course, to meet the constraint, the level of complexity featured by the law will have to be high enough to exclude many very complex systems. Additionally, the completed laws would have to say much more. Such things as the structure of the field, its character, and the timing of its evolution will have to be systematically related to properties of the physical system. Any theory of consciousness should explain these things, though, so they are not special problems for this option.

Concepts employed. The antecedent of this law has a very unlikely form for a fundamental law of nature. The concepts of *system* and *complexity* are each too vague to govern a phenomenon as definite as the phenomenal experience of our minds. Sharpening and defining them will not be easy, at least not in a way respectful of their new status of characteristics that nature is sensitive to on a fundamental level. I leave aside the difficult task of defining *system* to concentrate on the problems associated with giving a definition of *complexity*.

Complexity comes in types. There are the structural complexities inherent in the spatial organization of a system's components, the functional complexities of a system's contribution to a larger system, the computational complexities associated with the range of internal states the system may evolve through, and the relational complexities exhibited between a system and its environment.

Complexity also varies in description-relative ways. For instance, is the activity of a cell, described in terms of the molecular and atomic interactions within it, more or less complex than the activity of a brain, described in terms of the interactions of cell assemblies? Should we include the complexity of the cells in our account of the complexity of brain functioning, or should we abstract away from cells and treat them as primitive functional units?

How different complexities are measured can vary along dimensions internal to the kind of complexity being considered, and choosing the appropriate dimensions presents problems. Consider the problem of quantifying the complexity of the Amazon rain forest. Which dimensions should we pick to measure it? How do we find the "units of measurement" that would quantify it? And how should we quantify complexity to allow comparisons across radically different kinds of systems? Economies, brains, weather systems, and ecologies spring to mind as examples. Which is more complex and why: Einstein's brain or the system of global ocean currents?

Because of these kinds of issues, it does not seem to me that "complexity" is the right kind of concept to feature in a fundamental law of nature. It is too vague and varied and does not clearly refer to a natural kind. Also, these problems with complexity exist just as strongly for the concept of "system," which makes the proposal that complexity is the key ingredient doubly troubled. The theoretical efforts needed to address these issues are great, whereas the likelihood that the payoff will be a concept able to do both jobs of featuring in a fundamental law and restricting consciousness only to cognitive systems seems low. This makes it an unattractive direction of inquiry.

Simplicity. The moral of the preceding discussion is that the terms occurring in the antecedents of fundamental laws of nature must be sharp, not vague. Furthermore, articulation of a law relating qualitative fields to complexity will require identifying a favored kind (or kinds) of complexity (computational complexity, perhaps) and proposing that nature must be uniquely sensitive to this kind, as it alone gives rise to consciousness. Whatever specification we give will count as a new fundamental feature of nature, because nothing like, for example, "computational complexity" currently features in any fundamental laws.

Everywhere that the identified kind of "computational complexity" (for instance) shows in our fundamental law, it will of course be read as shorthand for its definition. Furthermore, to avoid panexperientialism, the analysis of complexity will have to yield a metric fine-grained enough that it can differentiate between the specific kind of complexity relevant to a cognitive system and the complexity in related noncognitive systems such as the earth's weather, an

economy, or a forest of trees. That means that the whole analysis of a system's complexity, added to nature as a fundamental feature, will figure into the application of the resulting law. This almost guarantees that the fundamental laws governing consciousness will be exceedingly complex, so believing that the laws are really fundamental will be difficult.

Empirical questions. Even if we imagine biting the bullet and accepting this added messiness in nature, it is doubtful that we can specify a dimension of complexity that will give intuitively satisfying results. We will need to set an upper bound on the values of complexity that lead to experience arising, and it will likely raise implausible cases.

If we do not posit a law that sets an upper bound on complexity, then we will not be able to avoid the kind of panexperientialism that tying consciousness to minds is supposed to avoid. The problem will be that any system in which conscious components are involved likely will count as at least as complex as the conscious components themselves and so would have to be conscious.

Consider a basketball team executing a beautifully efficient fast break. Each player has to track the movements, timing, and geometry of relations between all the other players involved, coordinating that with his own movements, moving the total motion of the group precisely toward their shared goal, tracking and handling the ball, and controlling his whole body on the ground and in the air. Participating successfully in this kind of coordinated autonomy is an incredible cognitive achievement for each player, and for the system corresponding to the players' disciplined interactions to function it must encompass essentially all the cognitive capacities of *every* participant. Thus the system will be more complex than any player alone. According to the no-upper-bound theory of complexity, theirs would be a superconsciousness, having experiences distinct from any single player's consciousness.

The prospects for specifying such an upper bound are extremely dim. We can foresee how a given proposal for an upper bound, motivated by considerations such as the aforementioned, will always admit cases of single systems (e.g., a creature with a brain twice as complex as a typical human brain) that we would never want to exclude from the consciousness club. Eventually, tremendous problems will arise for those of us who wish to individuate all the consciousnesses that must exist, problems tied in with how we are going to define *system* as it occurs in the fundamental law.

6.3 Functionality

Mere complexity possesses obvious shortcomings as the basis for consciousness. When someone suggests complexity, the real motivation usually has something to do with complexity being necessary for the existence of something else, like some kind of functionality. The natural reaction to the previous discussion is to suggest that our new fundamental laws supplement complexity with this something else. The most promising strategy is to delimit some range of capabilities,

like those in Kirk's *basic package*, as being critical to the arising of conscious experience. This is the second option: Perhaps qualitative fields are associated only with systems that possess certain kinds of *cognitive functionality* suitably defined. The coarse-grained form of the proposed law would be: *If a system has paradigmatically cognitive capacities XYZ, then it will have an associated qualitative field coevolving with it.* This position is quite common, although proponents usually put it forward as a version of nonreductive functionalism. Recently, Tye (1996) and Lycan (1996) have taken this tack, and a precursor may be found in Van Gulick (1988). Kirk suggests his *basic package* view as a form of analytic functionalism rather than nonreductive functionalism.

Concepts. Upon hearing this kind of suggestion, we need to remind ourselves that we are searching for something fundamental in nature. Fundamental laws of nature govern the behavior of fundamental things. The kinds of laws we are looking for are on the same level as those governing gravitation, motion, and mass. Whatever concepts occur in the antecedent of these laws will impute direct causal relevance to the things that fall under the concepts, a causal relevance not derivable from any constituents. This is very different from the kind of causal relevance, typically derivative on the causal natures of constituents, enjoyed by other high-level phenomena. As anyone familiar with twentieth-century psychology and philosophy of mind is aware, including the kinds of concepts being proposed here into any laws governing the qualitative field will make those laws very complex and will impute a character to natural laws not at all in harmony with those we have already discovered.

The fundamental problem is that defining cognition requires at least a *prima facie* appeal to *norms* and it is extremely unclear that the reference to norms can be eliminated in a way that is compatible with the invocation of fundamental laws. Like the concept of complexity, the concept of cognition is a vague matter. Presumably a correct account will involve concepts such as "*appropriate* behavior," "*veridical* perception," "supports mental states like beliefs that can be *true or false,*" "is *rational,*" and so forth. Kirk's *basic package*, to take the example introduced earlier, appeals to the ideas that such systems have *goals, assess* information, and *control* their own behavior. These concepts are all *intentional*.

Cognitive systems, because they have intentionality, are the kinds of systems they are because they can make *mistakes*, and the possibility of error requires the existence of a *norm* (Millikan 1984; Dretske 1986). Imagine the problem in the context of a developing fetus. When does the activity originating at its sensory surface change from simple causation to the production of a perception? Presumably when the activity produces representations but, *prima facie*, it seems the defining difference between simple causation and the activation of a representation is that the latter can be accurate or inaccurate. Similarly, at what point would the fetus have goals, attain self-control, be subject to illusion or error, or make decisions? How would nature know? It would seem to require nature to evaluate the system's cognitive development, as it is functioning in the womb, against a

Platonic conception of ways such systems *should* function and further determine whether it is *rational* (or *appropriate*) to hold the system to the standard.

A proposal that seems on its surface to meet the challenge is that cognitive systems are just a subclass of computational systems: According to the computational theory of cognition, any system implementing a member from a defined set of algorithms would necessarily be a cognitive system. These implementations would meet the conditions for applying cognitive norms, but they would not, themselves, be described normatively. However, the proposal just pushes the problem up a level. Computations provide rules for moving through a series of states so as to map inputs to outputs and nothing more. Nothing internal to the story of computation makes this mapping correct or incorrect, rational or irrational, adaptive or maladaptive. Nothing internal to the computational story even makes it the case that the computation is an information processing state or has a semantics. So the computational story itself does not say what makes some computational systems *cognitive* so that we can specify the relevant subset.

On the surface it would seem that we had a descriptive account of the conditions necessary for generating a qualitative field. Yet at a deeper, implicit level, it will still seem as if nature is mysteriously honoring a set of norms. We will make definitional decisions to sharpen our vague normative concept of cognition when specifying the class of computations, and so we will need to invoke norms to explain the lines drawn through the space of possible computational systems. Also, that subspace of systems itself will seem arbitrary along its edges.

Simplicity. A worse problem arises with the question of how we will formulate the fundamental law. An infinite disjunction will not do because, even if such disjunctions can describe real higher level properties, they should not be part of real theories of fundamental properties. The only alternative is specifying a precise set of descriptive rules that are coextensive (or close enough) with our normative criteria for cognition. Even in the unlikely event that we can specify such rules, they will not be simple.

More deeply, although every *specific* cognitive system will have some descriptive features that account for why it is cognitive, we cannot assume that cognitive systems *in general*—systems initially identified because normative concepts apply to them—share a set of necessary and sufficient *descriptive* properties. Because we cannot assume that *cognitive system* picks out a natural kind, we cannot assume that a *general* analysis of cognition will be able to eliminate implicit or explicit reference to norms. Therefore, if we are to specify the class of functional systems via some precise analysis of what it means to be cognitive, we can see already that the antecedent of this law will be implausibly complex.

Empirical questions. What would it imply about our world if we found a law of nature that seemed to respect normative considerations? It begins to look like our fundamental law would be expanding the character of our world far beyond just adding a primitive phenomenal component to our psychology's ontology. It

would be adding, as a fundamental feature of nature, a power of semantic divination. This is immensely far from the character of the fundamental laws we have thus far discovered. We expect a fundamental law of nature to appeal to purely descriptive conditions attaining in the physical system. It would be quite a peculiar law of nature that took effect only when a system was capable of making mistakes.

Leaving aside the pessimism I express here, even if functionalism is true, we likely will still need teleology (Lycan 1987). Teleology is the identification of something by its purpose or future cause, and it is necessary here because the functional character of a given state cannot be determined locally in space or time. The functional contribution of a state is relative only to a series of other states, to interconnections between them, and against the background of *appropriate* environmental conditions that could provide inputs.

The factual existence of consciousness will always depend, then, on a vast number of counterfactual truths extending widely through space and time. These facts are of the form: *If the system is in state-type X, and were it to encounter input-type Y, then it would transit to state-type Z.* Here, the state-types and inputs themselves are supposed to be functionally or teleologically defined, and thus *their* identity determination requires reference to further counterfactuals, and onward in a grandly holistic fashion.

We know now that nature does exhibit such counterfactual sensitivities on the quantum scale (Penrose 1994), so it is not something entirely new. Yet the form of the law we are now considering will require such activity routinely and ubiquitously on unheard-of scales, in macrosystems not known to exhibit these quantum effects, and at a level of sophisticated sensitivity completely without precedent. A law of that kind cries out for further explanation because it shows the arbitrariness and ad hoc character we normally interpret as signs that something deeper in nature is being exhibited.

6.4 Biology

A third alternative ties consciousness to biology by specifying that only biological systems reaching a certain level of complexity (or capable of certain kinds of functionality) can be conscious. That would rule out (presumably) global weather patterns, economies, and ecologies. Searle (1992) and Block (1980) have defended positions something like this. Of course, this proposal will share the problems plaguing the first two proposals, and it also has problems of its own. Were it true, there would be a fundamental law of nature whose antecedent contains a clause to the effect that . . . *the system is of complexity N and is carbon based* . . . Or, worse, a disjunction of the form *the system functions cognitively and is carbon or silicon based with a cellular construction.* . . . Proponents are hoping, of course, that we will discover something subtle about cellular mechanics or chemical reactivity that, combined with complexity or cognition, make it uniquely relevant to consciousness. The coarse-grained forms given here are

therefore caricatures, admittedly, but they are caricatures that capture the essence of their subject.

Concepts employed. Making a satisfactory case for any kind of biological constraint will be difficult. Consider two recent lines of thought that have been popular in the philosophy of mind, the appeal to evolution and the appeal to "wetware." Ruth Millikan (1984) has argued that evolutionary considerations are essential to understanding the intentional properties of systems. Because consciousness seems rife with intentional content, one might want to suggest that evolutionary considerations are also important for understanding why consciousness arises.

Some people with views on representation that are similar to Millikan's views seem to want to do this (e.g., Dretske 1995; Lycan 1996), but the position is extraordinarily implausible. The main concern of Millikan's semantics is accounting for the normative aspect of content, and we have already seen the implausibility that the kinds of laws we are searching for involve normative considerations. The difference between consciousness and intentionality (as Millikan conceives the latter) is that consciousness is undoubtedly intrinsic, "in the head," and basic. Millikan argues at length that content does not meet any of these conditions. Therefore, the analogy between Millikan-brand intentionality and consciousness breaks down.[3]

A more plausible suggestion is that somehow the wetware of the brain has unique properties that produce consciousness. The best option here is that chemistry might allow for certain kinds of causal relations that solid state physics cannot allow for and that these causal relations support consciousness. However, the danger here is to prevent an appeal to unique kinds of causal relations from being a disguised way of restating the functionality requirement, adding to it the additional bet that only biology could implement the functional conditions (which is not so implausible itself; see Edelman 1989 for suggestive observations along these lines). But then we have not really advanced from the functional criteria.

Empirical questions. Likely, the biological markers for given experiences will not support the events underlying a *unique* kind of experience. Much more plausibly, at different times the same biological objects will support different kinds of experiences by participating in different kinds of events. Our theory is likely to have the consequence that the activity of a cell assembly (for example) can at times give rise to a little patch of phenomenal blue and sometimes a little shooting pain and sometimes a little moment of angst and that its participation in different overall states attaches these patches to it. If it turns out that a neuron's firing as part of a "blue qualia" event is not significantly biologically different from its firing as part of a "purple qualia" event, shouldn't we factor out the biology by giving a more general characterization of the event types? Good methodology would say that we should.

I believe we will inexorably find ourselves pushed to a view in which the

event types are doing the real explanatory work. They will be differentiable from each other holistically, as playing certain roles in the continuing evolution of the system as a whole, and they should be carrying information based on the history of the system. Furthermore, for the purposes of explaining experience, these events will almost certainly be functionally construable, because adding " . . . and the neural firings were realized in a system using axons sheathed in myelin, and chemical signals involving serotonin, dopamine, glucose, and . . ." will add nothing to differentiate one experience from any other experiences. If so, biology will be an explanatorily impotent *X*-factor in our explanation of consciousness.

This consequence would be avoidable only if we could not give an explanation of cognitive events that abstracts from the biological substrate. Should that be the case, however, it likely would show nothing more than the fact that the biological substrate is the only substrate capable of supporting the proper cognitive functioning. This is a real possibility, but if that is the reason biology is relevant, then the relevance of the biological conditions will once again be derivative upon the relevance of the functional conditions. In essence, biology will be important only because of its unique capacity for supporting certain kinds of functioning. In this case, the biological requirement becomes redundant, because the functional requirement implies it.

6.5 Summary

My discussion here does not *prove* that the fundamental laws fail to tie qualitative fields uniquely to cognitive systems. The point rests on convictions concerning the simplicity, clarity, objectivity, and elegance of fundamental laws. They are convincing only to the extent that one shares these convictions about nature. If one does, then these kinds of considerations can yield strong reasons for rejecting the cognition constraint. By rejecting it, we set an expectation that the basis for qualitative fields will place them more uniformly throughout nature. It follows that the pretheoretical probability is that cognition merely represents a specific context wherein a more ubiquitous natural basis for experience expresses itself. Given the problems with each of the likely criteria, our fundamental laws are likely to have panexperientialist consequences.

7

Paradoxes for Liberal Naturalism

7.1 Introduction

The preceding chapters revealed puzzles and tensions for Liberal Naturalism beyond those associated with orthodox psychology and neuroscience. Those tensions suggest that Liberal Naturalists might have to rethink nature quite generally, and the puzzles raising these tensions are *clues*. For example, the Boundary Problem from chapter 4 points us toward something fundamental, such as causal interaction, in a search for conditions that create inherent individuals. Panexperientialism suggests that the conditions we are looking for exist throughout the natural world and take a specific form in creatures like us.

This chapter articulates five further issues for the Liberal Naturalist, each having the character of a paradox. For now, I am not making any commitments about where any errors might lie (although I make suggestions about a direction of inquiry to which the clues might point). I mostly want to expose the intuitions behind the paradoxes so that we might later diagnose the problems from *within* a Liberal Naturalist framework. I ask the reader to follow me in my restraint and resist the temptation to think quickly ahead, looking immediately to solutions, perhaps being prematurely tempted to take deflationary attitudes toward some or all of the problems.[1]

Liberal Naturalists have concluded that physicalism is false. Even though the arguments against physicalism rely only on minimally controversial observations about the nature of conscious experience, quickly dismissing more controversial claims about consciousness would be incautious. As Liberal Naturalists, or those considering Liberal Naturalism, we are no longer under any pressure from physicalism to embrace deflationary claims about consciousness. Were we to indulge in a quick dismissal of controversial claims on the grounds that these features of consciousness seem incompatible with the physical facts, we would be especially unjustified.

From a new perspective, we may be able to see that our situation does not warrant a deflationary attitude. After we enrich our view of nature, we may be able to resolve the paradoxes without convicting ourselves of hopeless phenomenological confusion or naiveté. At the very least, not *all* of the paradoxes rest on obvious errors or naiveté. By the end of this work, I will be in position to argue that we have been overlooking some possibilities concerning what nature is like, and these possibilities will provide a license to treat each puzzle with deference.

7.2 Category 1: The-Many-That-Are-Yet-One

1. *The unity of consciousness.* My visual field right now is teeming with phenomenal information. It represents depth, color, shape, motion, and, at a higher conceptual level, saliency. From an external perspective, one might think that each piece of phenomenal information could be present in a separate phenomenal modality. In reality, the coherence of presentation, in some sense, transcends the separation of content. These pieces of information have coalesced into what seems a unified field of perception, each piece superposed in an orderly way with the others. Even the separations between different sensory modalities seem superposed, in a subtler way, with one another in a common field. The remarkable character of this coalescence almost forces an inchoate belief in something we feel inclined to call the unity of consciousness. Thomas Metzinger (1995) attempts to describe it this way:

> I think that there is a highest-order phenomenal property corresponding to this classical concept of "indivisibility": The property of wholeness. The wholeness of our reality (and of ourselves in it) can easily be discovered by all of us from our own experience. This wholeness is much more than a simple unity in the sense of the concept of numerical identity mentioned above: I am not able voluntarily to split or dissolve my global experiential space—*this reality*—or my own experienced identity—*myself.* (p. 426)

The unity of consciousness presents both a paradox and a challenge. The challenge of unity is this: What do we mean when we refer to the unity of experience, and why *exactly* is it problematic? The challenge of unity is exemplified by Metzinger's decision to bypass description in lieu of an appeal to what "can easily be discovered" from experience. Among the puzzles and paradoxes of consciousness, a clear articulation of what we mean by the subjective unity of consciousness easily hides the most elusive description of all.

To say that its exact articulation is elusive is not to say that no definite intuition is there or that it is inarticulable. For instance, some people point to the nonlocality of quantum coherence, claiming that it is evidence for consciousness in nature (e.g., Kafatos and Nadeau 1990). At first glance, it might seem difficult to see why quantum nonlocality should make the existence of consciousness in the world more intelligible to some people. On reflection, I think the inseparability of the components in these entangled states resonates with certain prior intuitions people carry about the unity of consciousness.

I believe resonance with similar prior intuitions was an important initial motivation in the search for a solution to the *binding problem* for percepts. The binding problem is a problem about how the different information presented within a percept achieves the unified phenomenal character that we find in experience. As I noted in the opening paragraph of this section, the visual information in the separate processing channels for color, shape, depth, and movement all seem superposed into a single visual representation of the world. Translated into cognitive terms, the binding problem requires understanding how the information in these separate channels functionally comes together.

Prior to deep investigation of brain activity, external evidence did not make the binding problem compelling. From such a perspective, one could make a prima facie case that it is a pseudo-problem. Externally, the absence of any detailed understanding of the architectural constraints needed to produce human performance leaves us short of being able to justify believing that the brain globally binds separate information streams to achieve its results.[2] Whatever promising results the science is beginning to show, it seems the phenomenology of perception was the original basis of the belief. Our base belief in the binding problem originates from the conviction that something in the story of perceptual processing, when finally fully developed, should resonate with the phenomenology of inseparable, coherently superposed perceptual elements. Somehow, the idea of coherent neural oscillations, rather than independent or semi-independent processes, coheres better with our intuitions about the unity of experience as an explanation of perception. In short, something about causal inseparability and functional coherence remind us of the unity of consciousness. I suggest that beliefs about this inseparability of part from whole underlie intuitions concerning the unity of experience.

If that is the challenge of unity, the paradox of unity is this: We have unity of experience despite being composed of many diverse and independent parts. Conscious states are clearly complex, involving a tremendous variety of qualia coexisting within the experience of the subject. I suggest the unity intuition springs from the observation that, although the phenomenal manifold *is* clearly complex, it is *not* clearly composite. If it were a *composite system*, it would be a complex system whose existence derives from relations between independently existing elements, its components. The relation between a composite and its components is the standard bricks-in-the-wall one in which the existence of the components and their relations constitutes the existence of the composite. The components do not presuppose the existence of the whole in turn.

Composites and their components seem to provide the wrong model for understanding the relationship between our complex phenomenology and its elements. Although a clear distinction exists between the feelings involved in orgasms and headaches, it is not plausible that these feelings can exist independently of the whole experiencing of the subject.[3] Similarly, visual experiences are not plausibly constructed from tiny colored dots that exist independently of the experience as a whole, getting drafted into service in its construction. The intrin-

sic distinctions between these elements of experience do mark them as different, making for a complex and richly structured phenomenology. Their holistic dependence means that, despite being distinct from one another, they are not components of a composite whole. Instead, the elements of a given experiencing come into existence together, each dependent on the existence of the whole and strongly inseparable from it.

Our ordinary physical image of the world could not be more different from this, at least at the classical level. The brain is a complex system, and it is a composite system with components, too. It consists immediately of separable neurons and ultimately of fundamental particles. These particles support many layers of higher organization. Each of these higher layers of organization has individuals that are components in even higher levels. Eventually, we hit the biological level of organic molecules, nerve cells, neurons, neural assemblies, superassemblies, the brain, and the central nervous system in its entirety. The individuals produced at each stage seem strictly separable from the systems in which they are components (for an accessible discussion of the levels of nature, see Scott 1995). I can articulate both the paradox and the challenge of the unity of consciousness in these terms:

> *The paradox of unity:* How can a single system be both composite and noncomposite?
>
> *The challenge of unity:* A good theory should enable a more precise and clear articulation of what the unity of consciousness is. It should give voice to our unarticulated intuitions in such a way that we recognize them.

I believe current philosophy of mind may have a head start on solving the paradox. Notice that a *functional* system, considered *as* a functional system, has no component parts, although it is complex. Functional systems have functional entities as elements, and functional entities are what they are relative to the role they play within the larger system. Causal role typing is holistic, and any characterization of a type of causal role contains an essential reference to the role's operational context. These operational contexts constitute implicit reference systems against which the type of the functional entity can be defined.

Unfortunately, from a physicalist perspective, the kind of causal role typing at issue is just a conceptual exercise. Nature does not know about the existence of implicit reference systems against which one can define a functional role for things. The true physical reality contains no intrinsically favored contexts or roles. However, Liberal Naturalists, freed from reductionism, may consider the ontology of functional entities in a purer way. Perhaps nature is able to provide a favored context for some causal roles *screened off* from physical being in some way? Although I have argued that functional explanation cannot provide an explanation of consciousness by itself, perhaps it may still play an essential role in solving certain problems.

A good solution to the problem along these lines would define "causal role" in a way that avoids objectionable functional teleology, interest relativity, and norms

and that explains how nature could be sensitive to them through a normal mechanism not jury-rigged to work only in cognitive contexts. A natural research program that grows out of these issues is one that seeks to understand causation. What is it to have a causal role, and what is it to be a canonical context for a causal role? As Liberal Naturalists, we should try to understand these issues more completely.

2. *The subjective instant.* Another paradox arises from the apparent simultaneity of experiences within consciousness. Critics might call this, with William James, the specious subjective instant, but we have agreed to lay aside criticism for the time being. The paradox is that we have simultaneity of experience, although temporally asynchronous brain events seem to correspond to those experiences. In some sense, we can understand the simultaneity of consciousness as the temporal analog of the unity of consciousness.

Subjectively, the elements of our conscious experiencing—sight, smell, sound, proprioceptive monitoring, autonomic monitoring, –and so forth—all seem to occur concurrently, in a seeming stream of subjective moments. From the perspective of the subject of experience, their concurrent occurrence is an objective and special truth for which no other observer exists and over which no other point of view can have the same authority.

However, it also seems that brain events cause, or correlate with, the stream of conscious experiences. The brain is a complex physical system extended in space and time, and the laws of special relativity apply to it. For a set of brain events, there is no privileged reference frame from which we can say with authority that they simultaneously occur, no matter how finely we localize the physical correlate of the experience. Thus no objectively specifiable set of brain events exists that are all occurring at the "same time." These relativistic asynchronous events seem to be responsible for our subjective experience despite experience seeming to contain absolute synchronous elements.

In short, the problem is not the timing of the physical and phenomenal events per se. It is that, when it comes to the brain as physical object, no privileged reference frame exists from which one can say its events are or are not occurring simultaneously. Yet, when it comes to conscious experiencing, a privileged reference frame does exist. From an introspective viewpoint, an experiencing of different modalities—for example, sight, sound, and touch—in which the experiencing of the different elements would be occurring nonconcurrently can seem inconceivable. But relativity tells us that there are no privileged reference frames. Thus it seems that something exists to which relativistic spacetime both does and does not apply.

The usual response to this paradox is to call into question the phenomenology of the subjective instant. For example, it is tempting to see Daniel Dennett's *Consciousness Explained* (1991) as a book-length response to the paradox, using doubts about it to explain consciousness away. Dennett offers an account in which the simultaneity of the experience is the result of a narrative rationaliza-

tion by our reporting mechanisms. A probable mechanism is one whose job would be constructing a representation that represents them—falsely—as occurring simultaneously.

For a Liberal Naturalist, one unsatisfying aspect of this approach is that it seems to beg certain questions by presuming that the physical account of time is complete. The Liberal Naturalist is not under the same physicalist constraint as Dennett and is free to consider the possibility that our understanding of time is incomplete in some subtle way. Some presumptive evidence exists that this is the case. That evidence rests in the well-known problems physics has explaining the direction of time and its seeming "flow." Perhaps we may eventually explain the direction and flow of time using only resources within physical theory, but, until that day, the Liberal Naturalist should be willing to consider the possibility that something outside physical theory fixes the facts about the direction of its flow. Along these lines, it seems plausible that the direction of causation might fix the direction of time, and it is not completely clear that physics tells the complete story about the process of causation. If this were true, a promising place for the Liberal Naturalist to begin searching for a resolution to the paradox would be in a theory of causation and its relation to temporal flow.

Summary. Together, these two problems constitute what Nagel (1986) calls *the combinatorial problem.* Nagel feels that the combinatorial problem and panpsychism are the two biggest hurdles for a dual-aspect theory of the mind. I have already argued that current orthodoxy miscasts panpsychism as a problem. Nagel may also be overstating his case about the centrality of the combinatorial problem, but it does suggest that some common sense assumptions we have about either the physical, the phenomenal, or both may be deeply wrong.

7.3 Category 2: The Paradoxes of Epiphenomenalism

3. *The knowledge paradox.* Sidney Shoemaker introduced the first epiphenomenalism paradox succinctly in his 1975 article, "Functionalism and Qualia." He wrote:

> To hold that it is logically possible (or, worse, nomologically possible) that a state lacking qualitative character should be functionally identical to a state having qualitative character is to make qualitative character irrelevant both to what we can take ourselves to know in knowing about the mental states of others, and also to what we can take ourselves to know in knowing about our own mental states.

Shoemaker is worried that, if functionalism is false (and certainly if physicalism is false), the relations between brain states and conscious states will be *accidental* in that the qualia involved in consciousness would make no contribution to determining our brain states. Because our brain states drive our behavior, including our knowledge *claims*, it seems that qualia would be irrelevant to what we could or could not claim to know.

The following argument produces the paradox. By hypothesis, the world is

causally closed under physics. This means that every event in the world that has a sufficient cause has a physical sufficient cause and any probabilistic causes that are active are also physical.[4] Observe that our brain states are the gateways to producing our knowledge *claims*, that it seems very implausible for our knowledge claims to be justified if the objects of our knowledge do not matter to the production of our claims, and that we sometimes make knowledge claims about consciousness.

Given that we are capable of making knowledge claims about consciousness, we need to understand how consciousness could be relevant to the production of those claims. To connect consciousness to the production of our claims about it, somewhere in our explanation of our knowledge we will need to appeal to the effects of consciousness on brain states. Now these brain states are solidly physical, and we are assuming the causal closure of the physical, meaning that nothing nonphysical can make a causal difference. But if consciousness cannot affect brain states, it cannot play any part in producing our claims about it, and so it seems that we could not really know about consciousness. Yet we do know about it. Hence, Liberal Naturalism is caught in a paradox.

I am stating the intuitive problem. The Liberal Naturalist seems committed to conceding that consciousness makes no contribution to the fact that we make the claims about it that we do, and that is deeply troubling. Because any accuracy in our claims about it would seemingly be based on fortuitous coincidence, it seems impossible that we could know about it.

This is an almost unbearably subtle problem. The two most obvious moves do not obviously succeed. One temptation is to just deny the first premise, concluding that the world is not causally closed under physics. Most modern philosophers and scientists are justifiably wary of the interactionist dualism it implies. It *might* work if causal gaps exist,[5] but there is no clear evidence that such gaps exist, and the reply is not convincing without scientific support.

A second proposal (see Jackson 1982) is that natural laws assure that we make correct reports of our experiences by creating some appropriate parallelism between consciousness and associated brain states. Imagine that the brain states the laws operate on are the same ones that eventually lead to claims about experiences and that the laws assure that the associated brain states reliably produce the experiential states that make the claims true. So, usually, when I say that I am having an experience of phenomenal red, I actually am having an experience of phenomenal red. My brain is producing one via this law.

Even if we put aside its obvious ad hoc character, the lawfulness proposal seems too weak to impart justification: A lawful parallelism between a claim about an event and that event is not enough to justify the claim. To see the problem, consider Trey, a young man who is lawfully connected to Java, a volcano on Mars. Imagine a very strange law, one assuring that whenever Trey thinks about Mars, Java erupts. Therefore, Trey's brain states are reliably correlated with Java's eruptions.

Now imagine that Trey was reading a long philosophical work on conscious-

ness and that he reached a point in the work that discussed this bizarre scenario. Impulsively, Trey convinced himself that it must be true of him. Through this circuitous route, Trey has come to believe that a volcano on Mars erupts whenever he thinks about the planet. He therefore periodically claims, *Right now, a volcano is erupting on Mars* as he thinks of the planet.

Trey's claims are true, and they are lawfully correlated with the events that make them true. Are Trey's claims justified? It seems that Trey may come to believe in the connection for completely unjustified reasons, and therefore those claims would not be justified. If he were making his claims for the reasons given here, for example, they would not be justified. Their truth would be luck, despite the reliable nomic connection between them and the volcanic eruptions. Trey is really justified in his claims only if he is justified in believing in the lawful connection in the first place. In short, justification is more than lawful correlation.

So we cannot deny the causal closure of physics without great difficulty, nor attribute knowledge of consciousness merely to reliable correlations between brain states and conscious states. It looks as if the truth about experience should be completely *hidden* from our cognitive psychology. How can we explain the peculiar familiarity we have with our own feelings?

What we need is a way of explaining the intimacy between the phenomenal subject and its physical states. The ultimate expression of this paradox is the conclusion that the Liberal Naturalist is committed to epiphenomenalism or interactionism, neither of which is acceptable because they rule in only the most implausible external connections between consciousness and behavior. Despite its gravity, there is room to doubt whether this is a genuine dilemma. Although it seems like a problem about consciousness on the surface, it may really be a problem about how we understand causation and its place relative to physical science. To fight it, the Liberal Naturalist may need a deeper understanding of causation itself. Perhaps a full analysis of causation will yield a place for consciousness that is neither epiphenomenal nor interactive.

4. *The superfluity of consciousness.* Assume that we reject interactionism. This leaves us with an apparent radical epiphenomenalism about consciousness, that which is central to our mental lives and sense of self. This kind of revelation about nature is very problematic because it means that nature is not parsimonious. An epiphenomenal consciousness represents a fundamental promiscuity in nature and so conflicts with the convictions needed by a thoughtful realist.

Realists about scientific theories believe that those theories convey to us what is actually in the world outside us. Realists can be contrasted with antirealists, who believe that scientific theories are just tools to help us make predictions about experiments but don't necessarily show us what the world is like beyond the outcomes of those experiments. A critical tool in theory construction is Occam's razor, which states that simpler theories should be preferred over more complicated ones, other things being equal. Unlike antirealists, realists must take Occam's razor to be a statement of faith in the parsimony of nature. Because re-

alism walks so closely to the razor's keen edge, our confidence in the approximate truth of our scientific theories is always in danger of fatally cutting itself. However, the existence of consciousness, if it is truly a superfluous epiphenomenon, shows that at least one wholly superfluous set of properties exists in nature. It would be a clear counterexample to the realist's faith. Consequently, we should be suspicious that nature may absolutely abound with superfluous properties.

If consciousness were superfluous, we would not have grounds to resist the idea that nature is profligate. If it has noneffective entities, it certainly may have many efficacious entities that are, from a theoretical standpoint, formally superfluous. Can the theories that we use to describe the world be radically wrong precisely because they are the simplest theories? Perhaps God is politically correct and tries to maximize diversity. The superfluity of consciousness would make Occam's razor seem much more like a pragmatic principle and less like a metaphysical tool. As such, it would undermine our confidence in both philosophical and scientific explanation to get at anything like ontological truth. The paradox is that we seem to have knowledge about nature's hidden truths even though an epiphenomenal consciousness provides evidence that we should not really take ourselves to have that kind of knowledge.

Summary The knowledge paradox is frequently cited, but the paradox of superfluity is not frequently acknowledged in the literature. Penrose (1989) and Hodgson (1991) take good shots at it, suggesting that consciousness gives the mind some nonalgorithmic power. However, even if a proposal like one of those is correct, it will remain unclear why performance of the proposed functions is linked in any way to the existence of qualitative feels. As the antiphysicalist has argued, any story about the functional role of consciousness will fail to be the whole story about it. To address these problems in the depth they deserve, I return to the question of epiphenomenalism in the next chapter.

7.4 The Grain Problem

5. *The shallow structure of qualia.* The fifth puzzle is the *grain problem*, as introduced by Wilfrid Sellars (1963a) and recently taken up again by William Lycan (1987)[6] and Michael Lockwood (1993). The grain problem comes from noting that the physical character of brain processing involves structure not possessed by phenomenal qualities. For instance, the structure of an expanse of phenomenally experienced color does not divide into finer and finer substructures corresponding to the microphysical structure of the brain or brain events. Occurrent phenomenal colors, such as blue, are structurally homogenous despite their physical correlates having a highly variegated structure. For a physicalist, the problem posed is to understand how a relatively homogenous quality can be identical with, or constituted by, a richly structured physical entity.

Sellars and Lycan both propose that our physical understanding of the world must be incomplete. Sellars challenges us to find the "nonparticulate" foundations underlying our particle-filled understanding of the world. Other physicalists sug-

gest that phenomenal qualities may have structure hidden from awareness after all (Van Gulick, 1993). This latter suggestion seems untestable, because any test would have to *assume* the materialist answer it is supposed to be investigating.

Liberal Naturalism is not challenged in quite the same way by the grain problem, but it does raise questions about what does determine the structure of a phenomenal property and how. Interestingly, by again appealing to the functional aspect of an entity, in isolation from its physical aspects, Liberal Naturalism could make some headway on this problem. The grain problem relies on the assumption that the *physical* character of the system is the ultimate basis of consciousness. For a physicalist maintaining that something physical constitutes every mental event, or is token identical to it, this truly does raise a problem. Physicalists are not free to ignore the presumed physical basis of a mental entity's existence.

Liberal Naturalists are free of the assumption that conscious states ultimately reduce to physical states at the token level. They may ask, What follows for the structure of conscious experience if we assume that it is the functional character of the system, and only the functional character, that is implicated in determining the character of consciousness? To see how attempts to answer this kind of question could be fruitful, consider a group of Finite State Automatons (FSA) connected to one another by a signaling system. Each FSA is defined by a series of states it may take, connected by transition rules. These states are the kinds of states they are because of their connection to other states within the network, not because of any internal structure they might have. For the purposes of the abstraction that defines their type, each state is essentially atomic. It is easy to imagine logically possible worlds with structureless particles implementing given FSA's: The particles can be imagined to have a state for each of the FSA states and can be imagined to move from state to state in a way that mirrors the FSA's transition rules. In general, no internal structure is essential for a given FSA's states to be the types of states they are, and none needs to be specified. Figure 7.1 depicts an FSA with two states, q_0 and q_1, and two inputs, A and B. It recognizes when it is input an even number of B's because it always ends in q_0 when that happens.

Information systems can be made from networks of FSA's connected by an appropriate signaling system. Within modern computational systems, entities called *objects* or *components* are individuated and treated as state machines similar in spirit to structureless FSA's. Figures 7.2 and 7.3 define an object state machine in a common visual language (UML) used for the task and interactions between state machines for a simplified business information system. In the interaction diagram of figure 7.3, the components are represented as vertical lines, and the messages between them are represented as horizontal lines. They exist in computational networks at a grain abstracted from lower level details, cycling through states and exchanging signals (called *messages* and *return values*) with other entities inside and outside the system. These components are functionally defined entities, and specifying one of them requires defining their interfaces to other en-

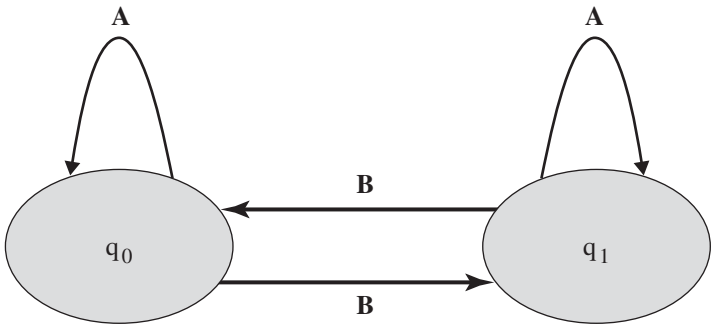


Figure 7.1 A Finite State Automata that accepts a string of *A*'s and *B*'s and calculates whether it contains an even number of *B*'s.

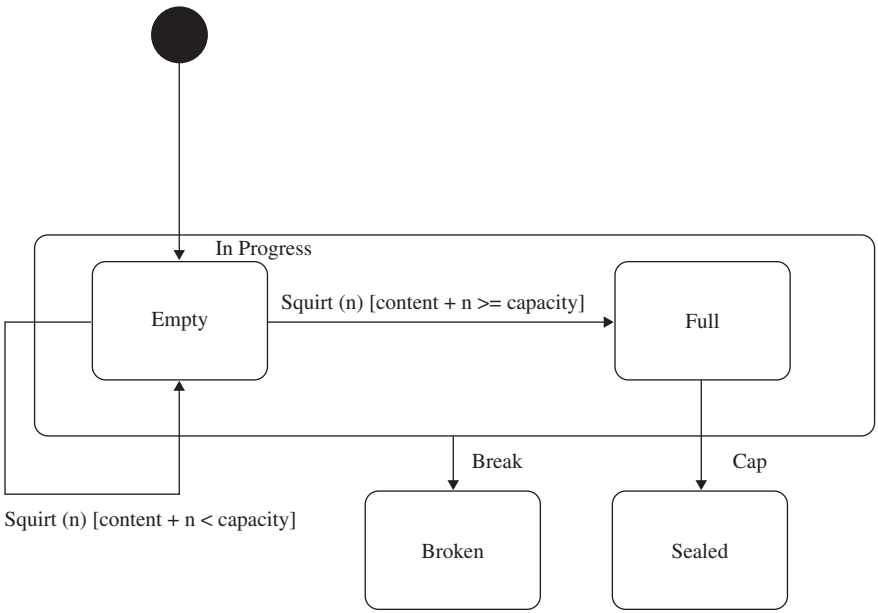


Figure 7.2 A state transition diagram in the software design notation UML. Notice that the machine states are structureless.

tities within the system, a set of signals passed through that interface, a canonical set of states for the entity, and rules for cycling through those states.

Components are purposely defined in a way that divides through the fine-grained details of their structure, both logical and physical, because there are usually many ways to realize the causal contribution they make to the functioning of the larger system. The abstraction from physical and logical detail of the states is achieved by allowing every entity to be wholly represented to other en-

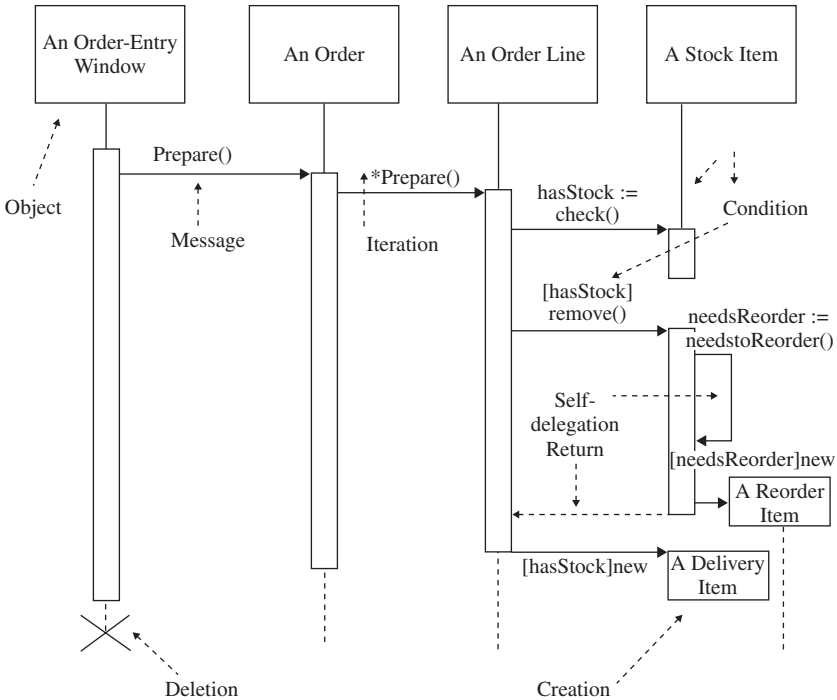


Figure 7.3 A sequence diagram in the software design language UML. It specifies messages passed back and forth between component machines, their order, and their effects.

tities by the way they participate in a shared signaling system (even behavior is defined in terms of signals sent and received). The signals a component uses to communicate within the system have a purely informational structure and are defined independently of the physical constitution of the system. Therefore, a functional entity's type and behavior will *always* be categorically independent of the fine-grained structure possessed by any particular realization.

To sum up: Functional objects are defined relative to a canonical reference system by the causal role they play within that system. These specifications of their causal role do not need to involve lower level structural details. Instead, causal roles are individuated relative to some level-encapsulated pattern of interactions. These interactions invoke other objects at the same level, are governed by a multidimensional signaling system, and, by relying on the semantics of the signaling system, implicitly invoke the referent systems they partially constitute. Ontologically, such entities subsist in the pattern of high-level differences their states can make to other systems at the same and higher levels through their participation in the signaling system. This means that objects and states are, qua functional beings, encapsulated at a level. As such, *functional Being has no finer grain*.

7.5 Summary

For a Liberal Naturalist, the prospect of solving the grain problem by appealing to the coarse-grained nature of functional Being is intriguing. As before, an appeal like that cannot take us the whole way, but it may play an essential part in a final explanation. Also, some detailed analysis of the terms involved must accompany it, terms such as *interaction*, *signal, level of nature*, and *canonical reference system*. In short, it requires investigating causal roles, which means understanding causation and the way causation individuates and stratifies entities in the world.

Every puzzle of consciousness can be construed as touching questions about causation in some way. This is the transition point of the book. In the next chapter I make this transition by setting up the problem of epiphenomenalism in a more rigorous way and then motivating a realist view of causation by developing some fundamental criticisms of conventionalist views.

PART II

Faces of Causation

This page intentionally left blank

8

Against Hume

8.1 The Specter of Epiphenomenalism

In this chapter I move from discussing consciousness to discussing causation. The threads of causation and consciousness eventually come back together in chapter 13, in which I argue that conscious experience is an aspect of causation itself.

Do physical explanations say everything there is to say about causation? Part I of this volume frequently pointed out how the nonphysical character of consciousness presents reasons and opportunities to look more closely at causation, especially causal interactions. One worrisome set of problems came directly from the causal relevance, or lack thereof, of consciousness. In their broadest forms, these worries imply that consciousness plays no role in the physical world's dynamical evolution. This is the problem of epiphenomenalism.

Epiphenomenalism seems to be true because the world's physical basis determines the dynamic and structural properties of everything and because people tend to believe that physics can tell us everything relevant about the behavior of the world's physical basis.[1] As a corollary, it is natural to assume that the *physical* explanation of our behavior is the complete *causal* explanation of our behavior. Here, *complete* has the sense that every entity relevant to a causal explanation either is physical or derives its causal relevance via its realization in a physical basis. This belief is called the *causal closure of the physical*.

Belief in the causal closure of the physical leads people to reject interactionist dualism. If one is convinced by the antiphysicalist arguments, then the most attractive remaining option is a dual-aspect theory. On reflection, the situation becomes paradoxical. It *seems* that we would say everything we say about consciousness even if we were not conscious. This includes all the pronouncements made by people like me to the effect that consciousness is unexplained and non-

physical. Why should a physical system ever accurately talk about its epiphenomenal, nonphysical aspects? Why should the physical processing reliably represent anything about an epiphenomenal, nomological dangler?

8.2 The Space of Possible Responses

Epiphenomenalist worries can make people wonder whether we might be deeply confused about consciousness, whether anything remotely like what we think of as subjective experience really exists. People are not convinced by the skeptical/eliminativist move primarily because our knowledge of consciousness does seem quite special. We are acquainted with consciousness. Because we do not postulate its existence to explain other things, failing to need it when explaining something does not justify radical skepticism. Yet resisting skepticism just puts us between a rock and a hard place, making it even more important to carefully think through the problem and answer it.

Epiphenomenalist worries are very serious, and they should inspire deep reflection about the premises that lead to them. The precise reasoning is:

1. The physical facts alone do not entail the facts about conscious experience.
2. We can conclude, from (1), that
   2′. Experience is a nonphysical aspect of the world.
3. A completed physical theory is, in principle, a descriptively adequate characterization[2] of the dynamical evolution of the physical world.
4. We can conclude, from (3), that
   4′. Our physical explanations are complete explanations of the causation involved in producing bodily movements.
5. We can conclude, from (2') and (4'), that
   5′. Consciousness lies outside the causal structure of the world, that is, it is an irrelevant epiphenomenon.

This argument creates a prism that refracts the different ways of placing consciousness in the natural order. Denial of premise (1) leads to reductionism or eliminativism, usually accompanied by attempts to show that our historic view of consciousness, the one taken from our everyday existence as conscious subjects, issues from a deep confusion. Defenders of eliminativism appeal to such ideas as that of self-monitoring robots that might talk about their cognitive states the way that we talk about consciousness. But our problem in explaining consciousness was never to give a physical account of what produces our *utterances*, as our utterances are not the grounds for our belief in consciousness. Awareness of experience itself is the grounding.

Denial of premise (2) leads to nonreductive physicalism, incorporating some kind of primitive metaphysical necessity. Appealing to metaphysical necessities does not help because the sense of necessity needed has never been specified well enough to make an appeal simultaneously effective and meaningful.[3]

Denial of premise (3) requires appealing either to interactionist dualism or to

brute emergence, along with downward causation. It faces the problem that it lacks empirical support. The spirit of the view is also thought by many to be in conflict with everything else we know about how nature works.[4] Considering these problems, a bet on one of these views seems like a bet on a long shot.

Finally, some people accept all the steps, along with the conclusion (5'). This leads to a kind of parallelism view, maybe supplemented with one-way causation from the physical to the phenomenal. It suffers from extreme counterintuitiveness and the air of paradox discussed earlier.

The one strategy that philosophers have not explored well is a denial of premise (4), where (4) is the inference from the adequacy of physical theory to conclusions about the causal completeness of physical explanation. This part of the book explores the possibility of denying premise (4). Making a plausible denial of (4) requires undertaking a detailed naturalistic analysis of just what causation is and the relation our physical theories bear to causation.

8.3 Problems with Hume's View

An initial reaction to the preceding argument might be a deflationary reaction appealing to a Humean view of causation. In Hume's view, the evolution of the universe is objectively unconstrained, and our causal stories are interpretive projections of the mind, a kind of psychological habit. In the *world* no connections of dependency, constraint, or production hold between individuals or events. According to this view, nonphysical mental events would have just as much right to claim causal responsibility for our actions as physical events, due to the regularities that hold between the mental and the physical. The Humean view has appealed to the antimetaphysical preferences of empiricists[5] for a long time because objective causal connections have seemed epistemically obscure and suspiciously metaphysical.

It is good to begin with this possible Humean response to the preceding paradox because the kind of substantial analysis of causation I ultimately develop contains natural ontology proposed on philosophical grounds and broaches metaphysics at many points. As a rule, global speculation and reorganization of this type is a response to equally global, categorical failures of explanation. The antiphysicalist arguments defended earlier and the associated paradoxes and puzzles that followed already establish this kind of categorical failure regarding consciousness. By rejecting Hume, I hope also to make the scope of the theoretical problems surrounding causation urgently felt. Specifically, Hume's view rejects exactly the element I choose to explore, the causal *connections* between distinct individuals and events.

After a long period of prominence, Hume's regularity view (which I also refer to as the *conventionalist* view) has begun to fall into disrepute, even among empiricists. Defenders of the view have never found a truly satisfactory account of what distinguishes certain regularities as being causal. Also, several insightful critiques have emerged to argue that the view has many other substantial shortcomings.

For example, Armstrong (1983) raises serious problems for the regularity view of natural laws. Problems that laws of nature pose for regularity views also undermine regularity accounts of causation, because motivating a regularity view about one without the other would be difficult and because we naturally expect that at least some instantiations of natural laws will be causal.

On the topic of explanation, Armstrong points out that, if the regularity view is correct, laws cannot be explanatory of the regularities they describe. Those regularities constitute the law, so citing the law in an explanation presupposes just what we are supposed to explain. Therefore, the regularity view cannot account for the explanatory role laws play in our practices.

On the issue of confirmation, Armstrong points out that the regularity view is susceptible to paradox. Because the basic regularity view is expressed using material implications of the form $(\forall x)\ (F(x) \Rightarrow G(x))$,[6] it is confirmed by instances in which $F(x)$ is false and $G(x)$ is true. For example, if it is a law that rising prices decrease demand, and if conventionalism were true, then this law would be confirmed by instances in which prices do not rise yet demand falls anyway (perhaps due to changes in quality, or demographics, or simply random chance).[7]

Armstrong also points out that the regularity view is not compatible with an understanding of laws of nature as real things independent of human convention. The problem is that any sequence of regularities in nature could potentially be explained by more than one hypothesis about an underlying probability distribution or deterministic function. Because the natural regularities do not uniquely determine the laws, neither probabilistic laws nor laws expressed by functions of varying magnitudes are entailed by the regularities they are supposed to describe. The Humean view must move very firmly away from being a reduction of laws to regularities and toward antirealism or irrealism about scientific truth.

Others have challenged the basic Humean point that we can coherently imagine any arbitrary relations of cause and effect holding between things and events. Harré and Madden (1975) point out that many referring descriptions incorporate a notion of causal production into their meaning. As an example, Harré and Madden claim that Joe could not be John's *father* unless he played a certain causal role in *producing* John. This role is not merely having a place in a sequence. Although no contradiction is involved in imagining that *Joe* might not have had this causal role in producing John, nevertheless it *is* contradictory to imagine that someone could be *John's father* if he did not participate in causally producing John. The conceptual scheme that gives meaning to "father" incorporates the idea as a truth-condition.

They also argue that we can properly apply certain natural-kind terms only to things that have specified causal natures. For example, identifying a substance that seemed to be copper by many tests but that still produced some different effects would present us with important conceptual challenges. In the end, either we would decide it was not copper after all, or we would have to engage in an extensive revision to our conceptual scheme to account for the unusual effects. Both possible results evidence a conceptual necessity that presumes a necessity

in the nature of the thing. We self-consciously involve the pattern of effects that copper produces as part of the meaning of the term "copper."

The discovery of antiparticles is an example of the kind of classification response at issue. A positron, which is the antiparticle of the electron, is indiscernible from the electron except that it has certain effects and responses that only a positively charged particle should have. Within the conceptual scheme of modern quantum theory, that quality is enough to disqualify it from being an electron, and the response was to create a new class of particle, strongly suggesting that our tacit understanding of what it is to *be* an electron includes producing the appropriate kinds of effects and responses. Someone who asserts that a particle that behaves differently from an electron is, nevertheless, an electron is either being incoherent or expressing an aberrant meaning with the term. Delimiting exactly what kinds of deviation force reclassification is a complicated issue, but the point remains that it is partly a conceptual issue about how we should rationally apply the concept. The conceptual issue centers on questions about what causal powers should enter into the meanings of the terms.

Harré and Madden (1975) urge that these observations demonstrate the unsoundness of Hume's argument that, in principle, we can coherently imagine anything having any effect whatever. The possibility that we can coherently, arbitrarily vary the effects of a thing is description relative at best because some descriptions and names incorporate references to bundles of causal powers into their meanings. What we can do is associate arbitrary effects with surface appearances or imagine one thing suddenly replaced by another with an entirely different nature. This ability is a far cry from Hume's original claims about our epistemology.

These and other criticisms present serious problems for the Humean view. Many regularity theorists have responded by proposing more sophisticated formulations of the theory. The development of the regularity view has paralleled Ptolemy's astronomy, adding epicycle after epicycle to a poorly conceived theory.

To cut this cycle short, we need criticisms of Hume that will apply to any theory that denies real connections of production or constraint between what exists or occurs in the world. So, instead of becoming entangled in a discussion of the epicycles that Humean theorists have produced, I focus on some less well-discussed problems that I believe are both broader and deeper. To explain these problems, I need only the bare Humean assertion that events do not constrain one another. Under that single assumption, the following problems seem to apply equally well to any possible formulations of the Humean view. The first of these problems is metaphysical and the second is epistemological.

8.4 The Metaphysical Problem: The Unity of the World

The world is complex, consisting of many distinct things. These things can be meaningfully related and compared along a variety of natural dimensions: in time, in space, in mass, in motion, in duration, as objects of knowledge and investigation, and more. This is the unity of the world. What allows for the exis-

tence of a single world in which there can be many things, all in a clear sense part of it, all capable of being meaningfully compared and related to one another within it, and what could provide a natural condition placing everything either inside or outside of it?

The intuitive way to account for the unity of the world is to take the causal closure of events. For example, assume that the Big Bang was the first event. Everything causally descended from the Big Bang is in its causal closure. Everything within the closure is part of the same world, and anything that falls outside would be part of a different world.

Using causal closure in the preceding way is a sensible and simple way to draw the boundaries of the world and to account for its internal unity, but it is a procedure that works only if facts about the unity of the world do not already figure into facts about causation. Otherwise, the attempt to account for unity using causal closure is circular. Humean views propose that regularities within the world *constitute* causal facts, and so these views must already have a unified world before they can account for causal relations. They cannot use causal closure to account for the unity of the world.

Origin of the problem. The regularity theorist must postulate atomic events, and by hypothesis each atomic event is itself a complete entity insulated from influence or constraint by the occurrence of any other events. To see that it is possible for a collection of independent events to fail to be a world, consider a collection of causally separated dimensions, such as a set of parallel universes in a science fiction novel. We can coherently conceive of each separated world as possessing its own internal time dimension. In this kind of multiverse, each world's time dimension would sequence the events within it. Nevertheless, there would not need to be an overarching, transworld time sequencing events across worlds. Thus there would be no answer to questions about whether event X in world A occurred before or after event Y in world B.

In the limit, each world could contain only a single event, with an internal time dimension giving it duration. But there would still be no transworld time that ordered events with respect to one another across worlds. A Humean world, with its insulated events, could very well reduce to this kind of multiverse of small worlds: Each event instantiates an internal time dimension that gives it duration, but there does not need to be a common temporal framework within which they all exist. Nothing would order them relative to one another, so they need not form one world rather than many separated, single-event worlds.

Because this first step splinters the Humean world into a multitude of separate and internally complete shards, the regularity theorist is faced with the problem of *putting the world back together*. The Humean cannot rest with a logical conjunction of events, letting their world be hauntingly reminiscent of the world in Wittgenstein's *Tractatus,* made up of a logical conjunction of facts. Let's call this the Humpty Dumpty problem, because the Humean must explain why we do not have a Humpty Dumpty world.

A Humpty Dumpty world would be death for the Humean view because no causal relations can hold between events unless they belong to the same world, or even if they all belong to the same world but temporal relations fail to hold between them within that world. On pain of circularity, the conventionalist cannot appeal to the usual unity condition of causal closure. So the Humean needs a different explanation for why (1) all the atomic events belong to the same world and (2) how they achieve the appropriate temporal relations needed for the existence of causal relations.

Spacetime as a solution. For the Humean, the main alternative to a Humpty Dumpty world is the view that spacetime is a primitive four-dimensional structure with a sui generis kind of unity. Three of the dimensions are spatial dimensions, and time is the fourth. Together they form a kind of seamless sculpture that supports events and provides an inclusion condition for the world.

However, the regularity theorist needs to account for the direction of causal facts, and the temporal dimension of the four-dimensional structure *has no direction that exists independently of a specification of causality!* Consider the light cone associated with an event in relativity, pictured in figure 8.1. We can picture the light cone as a cylinder squeezed in the middle to a point representing the point within spacetime at which the event occurs. The two conic halves represent the past and future relative to the event.

Which half is the past, and which half is the future? The standard theory of relativity says that the future cone is the one in which the event in the frame of reference may have causal effects; the past cone is the one that contains the event's causal precedents. Apart from this specification of how causal influence propagates, nothing in the theory itself determines which half is the past and which is the future. On pain of circularity, this is not an interpretation open to

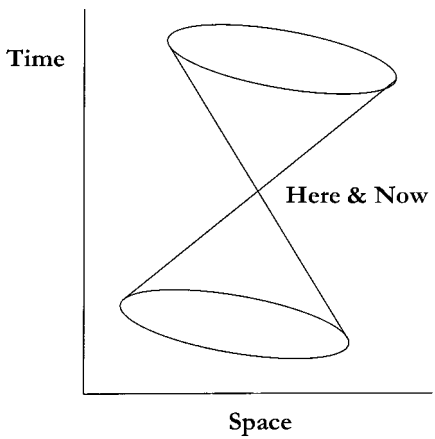


Figure 8.1 A representation of a relativistic light cone. Which half is the past and which is the future?

Humeans, leaving them in a world without temporal order and so without causal facts.

Subjectivism about time. Humeans are thus moved toward subjectivism about the direction of time, arguing that temporal direction is also a by-product of our human point of view. This is a position defended in a brilliant book by Huw Price (1996). For example, the second law of thermodynamics is asymmetric, as it states that entropy increases from one temporal extremity of the universe to the other. Might the second law provide recourse for a subjectivist about time, and through that to conventionalists about causation? Yes, but it does not come cheaply. As Price points out himself, the second law does not, by itself, give reason to believe that entropy increases toward the future any more than it does toward the past. It does this only in conjunction with the hypothesis that the universe had a special, low-entropy initial condition.

A conventionalist using the second law as an explanation for the seeming directionality of time therefore acknowledges some special facts about our universe: (1) the universe has a temporal extremity in a special and unlikely state; (2) only the *time* dimension—of the otherwise indistinguishable four spatiotemporal dimensions—shows a pattern of property instantiations that allows for lawful descriptions as one moves away from its extremity; (3) there is no sufficient reason why only one of the universe's four dimensions has this special feature; (4) one of these regularities (the second law) is asymmetric on the macroscopic scale; and (5) therefore, this special extremity is an *initial* condition. Put this way, the facts of the case seem incredible.

It seems to me that to acknowledge (1)-(5) is just to restate the mystery as its own solution, because the lawful distribution of events solely along the time dimension, including their conformance to the second law, *requires* explanation in a way that precludes it from *being* an explanation for the arrow of time. Events are lawfully describable when one follows their distribution from one extremity to another along the *time* dimension, but not when one follows their distribution along any of the *spatial* dimensions. Why not? What is so special about *time*? Conditions (1)-(5) set off time from the other dimensions of the spacetime structure without explaining why time should be so special.

If the second law explains the direction of time, then the kind of direction time has is *extrinsically achieved* by the arrangement of its contents, and it should be possible for any of the spatial dimensions to obtain the same kind of extrinsic direction by the arrangement of *their* contents. Imagine that we were traveling through space and came on a vast galaxy-sized region containing a great tumultuous cloud. This cloud is so vast that it spans billions of light years along its length. It is in great turmoil, its matter roiling and boiling and changing by the moment. Imagine that, just by chance, the cloud takes on a very unlikely but still possible state. Along the two-dimensional slice through the surface at its far end, the matter momentarily becomes highly ordered. Furthermore, the matter in the rest of the cloud is (by sheer coincidence from our perspective) also in a peculiar

arrangement relative to this highly ordered edge of the cloud. As one follows its volume along its remaining spatial gradient (imagine that we are traveling to our right along the surface of the cloud), the arrangement of matter along the two-dimensional cross-sections of the cloud is increasing in entropy. So the 2-D cross sections of the 3-D cloud are ordered along one dimension of space. Unfortunately, the cloud almost immediately passes through this state, and the order is lost forever.

Applying the time-subjectivist's principles to this cloud, the order of the matter within it would establish directionality for that region of that dimension of *space*, and this directionality would be *exactly the same* as the kind of directionality possessed by time. The cloud would have an initial condition and a future within space. Our cloud is just a 3-D analog to the larger 4-D universe, so the Humean has to agree here.

I raise the cloud example to ask a question. Given that space *can* be directed under the 4-D model, why isn't it? Equivalently, given that time could have its contents as *unlawfully* distributed as space does, why doesn't it? This fundamental asymmetry—that one dimension is lawful while the others are chaotic—is a striking difference between the spatial dimensions and time. It would be there even if all natural laws were time symmetric. Accordingly, appeals to the second law (or any specific law, it seems) do not answer the question. This question is natural and reasonable and demands a fair explanation.

Summary. Regularity theorists start by wanting to avoid making unintelligible metaphysical postulates about causal connections. To do that, they need to explain world identity and temporal precedence independently of causation. The most plausible way to meet the challenge is to reify spacetime, postulating the existence of a four-dimensional spacetime structure to house events. The contours of this structure seem to have no inherent *direction*. Therefore, the structure itself does not support facts of succession or precedence that can ground causal facts. In response, regularity theorists must simply assert that one dimension of this structure, and only one, has an extrinsic feature that yields a direction. However, this fact seems to be an extraneous adhesion to the model and raises questions about why the *temporal* dimension would have this property when the others do not.

8.5 The Epistemic Problem: Solipsism of the Present Moment

By the preceding argument, it seems that conventionalists are saddled with metaphysical claims more problematic than those they were trying to avoid in the first place. The quandary might tempt conventionalists into resisting the charge that their temporal postulate is ad hoc by saying that they know from experience that our world has direction in the time dimension.

This move would be ironic because it echoes non-Humean responses, rejected by Humeans, to Hume's own arguments about causation. If offered by a Humean here, it leads directly into an epistemic predicament that seems even more serious

than the metaphysical predicament. Hume himself argued that his views lead to inductive skepticism so that we cannot know whether the future will be like the past. Additionally, I argue in the following that Hume's view also leads to skepticism about whether the past was like the present and about whether external events are represented by internal events. In short, conventionalism simply cannot be saved from global skeptical consequences. I believe it reduces to solipsistic skepticism.

The origin of the problem. The epistemic problems exist because the Humean view reduces to something like a version of Leibniz's monad view. Leibniz's monads were absolutely unitary beings whose evolutions were completely independent and insulated from influence by other monads. Leibniz coined the term *windowless monad* to refer to this independence, the image being that every monad was absolutely closed to information about things other than itself. If the regularity view is correct, the world's basic events are separated, each occurring independently of and unconstrained by any other event. These insulated Humean events are excellent proxies for monads in a Humean world.

What happens to our conscious minds in this picture, with their collection of occurrent mental events? It seems that we must treat our conscious minds as monads or monad complexes. Applied to our minds, the conventionalist position implies that the mental events occurring in a subject are unconstrained by anything else in spacetime. These unconstrained mental events include perceptual events, occurrent thoughts, occurrent memories, and occurrent beliefs. In a Humean universe, which is a universe without Leibniz' beneficent God who guarantees synchronicity, a monad complex cannot have knowledge of the external world. The argument for this conclusion begins with a definition:

> Definition 8.1: A set of events *E* is *veiled* with respect to an epistemic agent *EA* if, and only if, *EA* is situated in the world in a way that prevents *EA* from gaining information about the events in *E*.
> Skepticism: If an epistemic agent is veiled with respect to a set of events *E*, then the agent cannot know about the events in *E*.

The argument is as follows:

1. Past events do not constrain future events. (Humean hypothesis)
2. If past events do not constrain future events, the future is veiled. (Humean inductive skepticism)
3. The future is veiled. (line 1, line 2)
4. *Future skepticism:* We cannot know the future. (line 3, *Skepticism*)
5. Future events do not constrain present events. (causality assumption)
6. Present events are the future of past events. (temporal logic)
7. Present events do not constrain past events. (line 5, line 6, substitution)
8. Past events do not constrain present events. (line 1 and line 6)
9. *Generalized inductive skepticism*: If (1) events in temporal region *T* do not constrain events in temporal region *T*,* (2) events in *T** do not con-

strain events in *T,* and (3) $T \neq T^*$, then epistemic agents in *T* are veiled with respect to events in *T**.

10. Past events are veiled. (line 7, line 8, generalized inductive skepticism)
11. *Past skepticism:* We cannot know the past. (line 10, *Skepticism*)
12. Perceptions are events occurring at a given time *T* that are about events occurring at a different location at time *T-k* for some measurable interval *k*.
13. *Perceptual skepticism:* Perception cannot give us knowledge of other events. (line 12, *Past skepticism*)
14. *Solipsism:* We cannot have perceptual knowledge, knowledge of the past, or knowledge of the future. The Humean view reduces to solipsism of the present moment. (line 4, line 11, line 13)

Humean inductive skepticism is a consequence of the Humean hypothesis that the past does not constrain the future. Given their hypothesized independence, even complete information about the regularities in the past does nothing to convey information about what will be the patterns of events in the future. Once inductive skepticism is recognized, generalized inductive skepticism follows for similar reasons, and global skepticism cannot be plausibly avoided. In a Humean world, any correlations between perceptions and external events or between memories and past events are coincidental. By eliminating the concept of constraint between events, the external regularities within a Humean world would seem to lose their relevance to what occurs within any given mind on any given occasion.

The epistemic situation of a believing subject in a Humean world is disturbingly like that of a person in a classic Gettier-type situation. Consider a stock Gettier example: A person looks at a broken clock, unaware that it is broken, and reads the time. The person has every reason to believe that the clock is working—that it tracks the time of day—and forms a justified belief that the time of day is as the clock says. Also, by pure coincidence, the person happened to look at the clock at precisely the time that is frozen on its face, meaning that his or her belief about the time of day happens to be true. So the person has a justified, true belief. The Gettier problem is that the belief is only coincidentally connected to the time of day. Because a lucky coincidence like that is no better than a blind guess, we still cannot say that the person really *knows* what time it is. If the Humean view is correct, our minds are like broken clocks that arbitrarily change their representations from minute to minute. In such a world, correlations among mental events and other events would be improbable and coincidental, and, because of this, would fall short of being knowledge.

Summary. I find the epistemic questions raised by Humean views to be very serious and unavoidable within that framework. Global skepticism looms as long as we must suppose that the mental events occurring within our minds are unconstrained by anything else occurring in spacetime. I can sum up the grounds for

skepticism by reiterating that mental events in the Humean world will not carry any *information* about the rest of the world, at least not in any sense of *information* strong enough to create knowledge.

The epistemic and metaphysical problems tie into one another. I started the discussion of the epistemic problem by supposing that our imagined regularity theorists propose that experience reveals a primitive direction of time and thereby deny the charge that postulating an unexplained direction to the temporal dimension of spacetime is ad hoc. The epistemic problem shows that we could not have information about the direction of time in our world if the Humeans are right. Specifically, if the direction of time is an extrinsic consequence of the direction of change in entropy, then we should be able to postulate coherently that it varies between different regions within the spatiotemporal structure. How do we know that we do not live in a world in which time is directed *this* way at some regions and *that* way at others? Actually, the problem is worse. The skepticism brought on by the problem affects memory, also. We would have no more reason to trust the deliverances of an occurrent memory than an occurrent perception. Subjectivity in the Humean world reduces to *solipsism of the present moment*. In a Humean world, we would just have *no idea* what was going on in spacetime.

8.6 Beyond Hume

The preceding discussion, if correct, shows the Humean view faltering on many fronts. It seems to fail at providing explanation, at accounting for scientific confirmation, and at avoiding dubious metaphysics; and it falls prey to skepticism of the worst kind. The arguments against Hume show that real causation, like consciousness, presents severe explanatory difficulties. We need a general and enlightening view of what kinds of properties a world with *those* kinds of facts—facts about connection and constraint—would require. What kinds of properties and structures must a world with real causation possess?

9

The Theory of Causal Significance

9.1 Introduction

In part I, I argued for several problems that face the Liberal Naturalist's program for explaining consciousness. These were:

1. A puzzle, called *the boundary problem for experiencing subjects*, about why conscious experience exists at the middle level of the natural world even though it seems coherent that things could have been otherwise.
2. The possibility of panexperientialism, a more benign form of panpsychism. It even seems *likely* to be the outcome of Liberal Naturalism.
3. The unity of consciousness as a property of a seemingly disunified brain.
4. The seeming existence of a subjective instant.
5. Problems associated with the causal relevance of any extraphysical aspects of reality.
6. Sellar's grain problem about the structural homogeneity of phenomenal properties.

While exploring most of these problems, I suggested ways to view them as providing reasons to look more deeply into causation. Discussion of the boundary problem ended with questions about how interactions might create layers of inherently individuated subregions of the world. The riddles surrounding the unity of consciousness and the grain problem could point to questions about causal-functional roles, and functional role questions are ultimately questions about causation and causal interaction. The paradox of the subjective instant leads to questions about time, and potential ties between the direction of causation and the direction of time are enticing targets for exploration. Most obviously, the problems associated with the causal relevance of consciousness cry out for an in-depth treatment of causation.

The argument that raises problems for the causal relevance of consciousness contains a promisingly questionable inference: It moves from the scientific adequacy of physical explanations to the conclusion that physical explanations tell everything fundamental there is to know about causation. To my knowledge, this inference has never been formally challenged. In this chapter I challenge that inference. To do so, I need to present a solid idea of what causation is and what a full explanation of causation should look like.

The theoretical framework I develop is called the *Theory of Natural Individuals*. The first piece of the framework, developed in the next three chapters, is the *Theory of Causal Significance*. This chapter is an introduction to the Theory of Causal Significance that is intended to motivate the general approach the theory represents and to introduce and explain the basic concepts. This chapter:

- Defines the problem of causation, explaining why a theory is needed and important.
- Explains why physics is not a theory of causation.
- Gives a taxonomy of traditional approaches to causation and explains why the Theory of Causal Significance must fall outside of the traditional taxonomy.
- Abstracts a very general essence of causation that the Theory of Causal Significance can represent and shows how to modify the traditional taxonomy to create a place for the Theory of Causal Significance.
- Emphasizes that causal significance is not necessarily the production relation of cause and effect.
- Introduces the ideas of effective and receptive properties, arguing that they are conceptually and empirically distinct aspects of causation. Together, these properties are said to provide the *nomic content* of an individual.
- Defends a proposal to treat receptivity as a connective property.
- Analyzes the causal nexus, defining key terms, giving examples, and laying down the fundamental principles of a theory of the causal nexus.
- Explains what a natural individual is and discusses how and why natural individuals might emerge at many levels of nature.

9.2 The Problem of Causation

What is the problem of causation? Imagine two great, blank canvases that you cover with color one drop at a time. Imagine also that the two canvases are very different kinds of surfaces with which to work. You call the first canvas the Humean canvas, and it will accept any drop of paint anywhere on its surface in any color that you let fall. If you let a drop of red paint fall onto the Humean canvas, it will stick where it lands. The same will happen if you then drop a speck of yellow paint somewhere else on the canvas. You can fill the whole can-

vas this way, dropping colorful spot after colorful spot on the Humean canvas until its surface is covered with colors lying beside one another in any combination whatsoever. The canvas cares not a whit what the end product looks like, ugly or beautiful or anything in between.

You call the second canvas the Canvas of Causation, and it is more of a marvel. If your first drop of paint is a bit of green, and then you try to place a dollop of red next to it, the red paint will bounce off. The canvas will not accept it. But it will accept yellow. And the more paint you put on the canvas, the more subtle and picky it becomes. Each bit of color that sticks to its surface seems to place a constraint on what colors may appear anywhere else on the canvas. In fact, although the canvas will allow you to paint it many different ways, it will accept only combinations of color that make for a beautifully covered canvas, so that somehow the canvas enforces aesthetic laws. Every color and every drop matters, jointly enforcing or excluding the colors that will finally appear on the canvas.

Although the Humean canvas is ordinary, the Canvas of Causation seems like magic. The two canvases are two possible ways the world could be. The drops of paint represent events that occur in the world, and the laissez-faire chaos of the Humean canvas represents a world in which anything can happen anywhere, regardless of what else might have occurred. The magical pickiness of the Canvas of Causation and its aesthetic laws represent a world in which laws of nature suggest a connection between each event so that every one must somehow respect the nature of every other. It is a world in which nature includes and excludes membership based on what else has made it into the club.

The problem of causation is that we do not live in a Humean world, even though the Humean canvas seems so much simpler to make than a Canvas of Causation. Making a Canvas of Causation requires some extra ingredient over and above simply having a world in which things can happen, and it is not clear what this extra ingredient is or what it means for our understanding of the world in general. Given that our world is like the Canvas of Causation, it seems that there is some magic in it somehow that connects things to one another in a deep way. The problem of causation is to understand what that really means for the nature of things.

9.3 Physics Is Not a Theory of Causation

On the path to understanding causation, the place to start is with physics, the aspect of causation that we understand best. A realist but Humean interpretation of physics is easily available to us, and this easy availability of a Humean interpretation exposes the danger that physics might not be telling us the whole story about causation. Physics might be describing only an aspect of causation, and, by realizing physic's potential shortcomings, we will be in a better position to find what is missing.

A description of coevolving fields is the centerpiece of quantum mechanics,

our most basic physical theory. These fields expand and periodically contract, for reasons still unknown, to something like classical, localized particles, and then they begin to spread in spacetime again. The dynamical laws tell us how any given field will evolve given its state at some time in the past, and they tell us how the evolutions of different fields become correlated. The current theory does have a gap in its dynamics because it must appeal to the ill-defined concept of measurement to specify when the contractions of the fields occur. This gap in the theory should not matter to the discussion that follows.

The evolution of a field is represented by a dynamical equation called the Schroedinger equation. Schroedinger equations plot states of the system, represented in a matrix, against points in time. Given an initial state, the mathematical rules they express describe a temporal trajectory through the field's space of possible states. The relevant feature of such dynamical equations is that their successful use requires us only to assume regularity in the succession of states. They merely associate, or correlate, field states with points in time. Association is a weak metaphysical relation because associations could exist for just about any reason or for no reason at all.

Specifically, the mathematical machinery is neutral with regard to how these associations arise. Nowhere does it mention or need the idea of causal production or dependency between states of the system at different times. The only explicit associations in the *function* are between states of the system and points in time. It is the explicit and implicit associations represented in the function that contain the causal content of the theory. There is no need for the hypothesis that one state of the system might causally depend on or be connected to another by more than their places in the overall extrinsic pattern. If we choose to interpret the mathematics causally anyway, this interpretation is projecting something into the theory not explicitly represented nor logically required by its equations.

The second component of physical theories describes how these fields "interact." I put "interact" in scare quotes because this part of the theory is also compatible with a Humean view of nature. The laws describing interactions express correlations between the evolutions of different fields. Like association, correlation is a weak relation and compatible with the absence of any real connection between the fields. It is true that physical forces are supposed to mediate these interactions, but virtual particles carry these forces. We can always interpret virtual particles as further field elements entering the correlation story.

In the end, a realist interpretation of the equations governing interaction requires only that we recognize the highly regular correlation between the evolutions of different fields. Like talk of connections of causal dependency, connections of interaction and exchange of "information" (in any active sense of "information") is projected into the theory. We do this because we would find the world the theory tells us about impossible to believe in without such connections and not because the theoretical apparatus logically requires us to think that way. Particularly, the theory does not represent causal connections. If we choose to interpret physical

theory in a non-Humean way, we must take it as assuming causal connections implicitly while explicitly describing some aspect of their outcomes. In this chapter and the next, I try to make the reasons for this clearer. In chapter 11 I give a formal argument for the conclusion.

One can think of this theoretical apparatus as a kind of probabilistic road map. It helps us navigate the four-dimensional surface of spacetime using landmarks to help fix our expectations. To be a successful map, it needs to make only modest demands on nature, not requiring anything more of nature beyond the regularity of relations between the landmarks.

The metaphor of a map tells us how we can be both realists and Humeans about physical theory. Corresponding to every physical property in the theory, we postulate something in nature. We can think of "mass," "charge," "spin," and so forth as each denoting a property present in the appropriate magnitudes at the appropriate places in spacetime. These properties are distinct and capable of the specified quantitative variations. They act as the landmarks on our maps. That makes us realists about the science because we are taking it to refer to objective properties belonging to things outside of us and describing them accurately.

The theory can be true, and true in a realist sense, even if we do not postulate further things such as connections between the landmarks. The landmarks simply have to vary in the regular ways that the theory describes so that spacetime has the appropriate layout. We do not need to suppose that some landmarks produce others or constrain the production of others. Therefore we will not postulate these things. That makes us Humeans.

Just because we can easily see how to be Humeans about physics does not mean that we *have* to be Humeans about it. Humean views have deep problems, as I argued in the last chapter, and the most common and compelling interpretations of physics are causal. I am suggesting that the ease and directness with which we can construct a Humean interpretation should serve as a warning that we cannot make the move to a fully causal interpretation for free. To make sense of unnoticed background assumptions, we may require ontology that physical theory does not explicitly represent. Perhaps we will have to take physical theories to be explicitly representing some *aspect* (or aspects) of causation, while allowing others to live *implicitly* in the background. The business of the Theory of Natural Individuals is to find and more explicitly characterize these implicit categorical grounds of causation.

Admittedly, I have not said anything about the hypothesized quantum collapse of the wave function or alternatives to standard quantum mechanics, such as hidden variable theories. None of these things make a difference to the general point, which rests on an observation about what our physical theories actually require from us to deliver their results. To do their empirical job of predicting or explaining what we observe at some region in spacetime, they require us only to possess certain minimal information about the values of physical properties involved in some other events that have occurred elsewhere in spacetime.

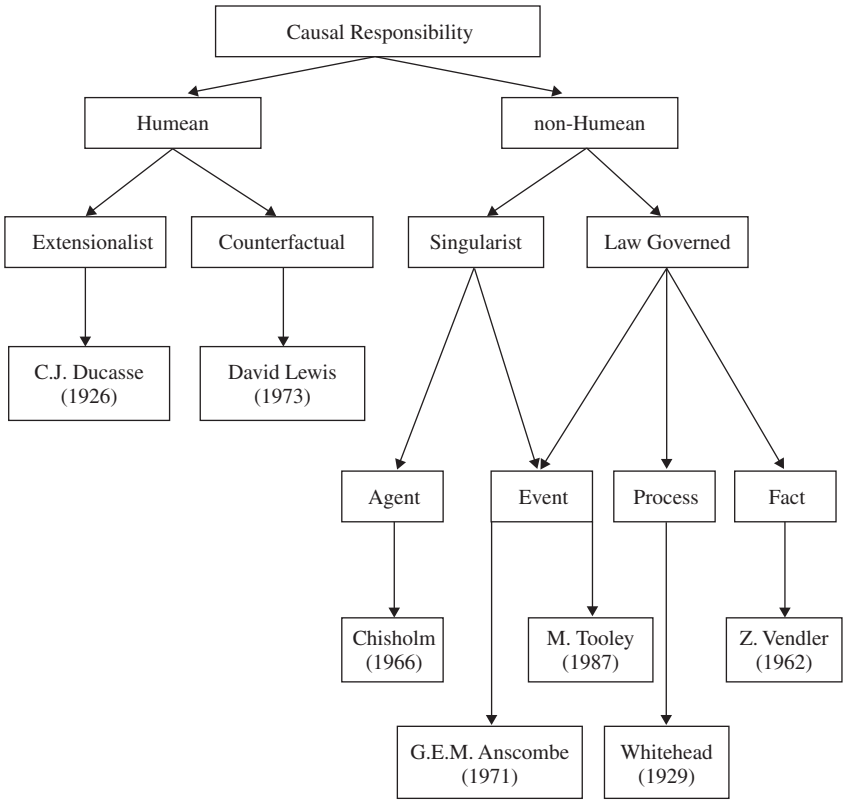


Figure 9.1 A taxonomy for traditional theories of causation. The leaf nodes represent chief proponents from the twentieth century for the corresponding view.

9.4 Causal Responsibility

Many philosophers of causation tacitly assume that their first choice is between a Humean conventionalist approach and some form of nonconventionalist approach. The decision tree that seems to be active among theorists is depicted in figure 9.1. The root node of the tree, labeled *causal responsibility*, represents the assumed ultimate object of explanation for a theory of causation. Facts about causal responsibility are facts about what makes a productive cause and how these causes relate to their effects.

Theories of causal responsibility are theories of general conditions under which a specified something—event, agent, fact, or process—can be credited with being the cause, partial or total, of some specified event(s), its effect(s). The branches of the tree are choice points along the way to developing this theory of causal responsibility. The approach I develop here differs from the standard approaches at this very first choice point by rejecting causal responsibility as the fundamental explanandum for a theory of causation.

I steer away from the tradition because it is not fully objective. Our ordinary notion of causal responsibility has strong intentional and interest-relative components. The intentional aspects betray themselves when negative facts show up as causes in both ordinary and scientific explanation. For example, when an animal starves to death, we judge that the cause of death, which is a *loss* of life, was the *lack* of food. Along the same lines, we often say that a person's disappointment in himself or another was caused by some *failure*, where failures are understood as things that were *not* achieved.

It is not so easy to eliminate the negatives from such examples and, more important, it is not worthwhile. For example, perhaps we may define death in positive terms as the presence of monotonically increasing entropy in the organism. Assume that a wicked pet owner starves his or her pet to death by locking it in a closet. What is the cause of the monotonically increasing entropy that eventually takes hold in the pet? There is certainly a complicated story concerning positive facts to be told, but this story is at a level below which we assign causal responsibility and misses the key fact: The pet was prevented from eating. A court would say the death was caused by neglect. Still, a coroner might cite liver failure as cause of death. A systems-oriented biologist might cite failure of systemic homeostasis. It depends on one's prior interests and point of view. Finally, for some negative facts, such as the feelings of disappointment caused by someone not showing up for a date, there truly seems to be no sufficient set of positive facts to substitute for purported negative causes. The problem raised by such scenarios is that facts about absence require appealing to intentional objects such as universal "That's all" facts. These are universally quantified facts that are logically equivalent to negative existential facts.

Furthermore, if one were to produce a complicated set of purely positive facts, assigning causal responsibility from this large set of positive facts yields to the problem of deciding what counts as figure and what counts as ground in such judgments. The interest-relative aspect of causal responsibility shows itself in judgments that essentially involve a kind of figure/ground relation. Imagine a typical morning when Trey goes to work. Before getting on the road, he puts the car key in its slot, turns it, and starts the engine. Although our common idea of causal responsibility will credit Trey's turning of the key as being the cause of the engine's starting, notice that the counterfactuals involved underdetermine this kind of judgment. Although it is true that the starting of the car would not have occurred had the key not been turned, this same counterfactual holds of many other facts: Had his morning alarm not gone off, Trey would still be sleeping and thus the starting of the car would not have occurred; had the spark plugs not fired, the car would not have started; had the earth stopped turning, the car would not have started; and so forth. The counterfactual seems to be an important condition, but the truth of such counterfactuals is not sufficient to yield facts about causal responsibility. Giving a sufficient account seems to bring in interest-relative factors relying on idiosyncrasies in human judgment (such as how we might judge the similarity relations between two possible worlds).

One might try to remove the figure/ground problem by expanding the scope of causal responsibility to include all facts necessary to produce the effect. However, we have learned now that previous states of the world do not necessitate subsequent states. Therefore, assignment of responsibility must come on some other grounds, such as making the subsequent states of the world more probable. Assume that time is continuous, and let *C* be a state of the world proposed as being causally responsible for an effect *E* occurring later in time. For any *C* and *E*, there will be a state of the world *C** between *C* and *E* such that *C** makes *E* at least as probable as *C* does and which is closer in time. There is therefore no objective reason—no reason which matters to nature—to make *C* rather than *C** the state which is causally responsible for *E*. The issue is decided based on human interests. Perhaps time is not continuous, so such problems are only apparent, but a theory of the deep structure of causation should not be hostage to such matters.

For such reasons as these, I believe that facts about causal responsibility are unlikely to be similar to facts about rocks, things that we simply trip over while investigating the world's objective causal structure. These aspects of our ordinary concept of causation create a striking portrait of a convenient explanatory construct rather than an objective natural relation, and judgments of cause and effect seem like ways of characterizing certain striking patterns. I believe these intentional and interest-relative aspects of causal responsibility are what can make the conventionalist views about causation seem plausible.

The intentional and interest-relative aspects of causation have been especially emphasized by R. G. Collingwood (1940). More recently, D. H. Mellor (1995) has emphasized the tight relation between the notion of cause and being a means to an end. To move past conventionalism, it will be necessary to dig through to an objective core. Because a metaphysically robust kind of causation must exist (per the arguments in the last chapter), facts about causal responsibility must arise from a mixed notion, one that contains an objective core on which the more intentional and interest-relative facts rest. We are stalking an explanation of this objective core, not causal responsibility itself.

9.5 Causal Significance

A robust metaphysical theory of causation will provide a viable realist alternative to conventionalism. The preeminent theoretical virtue guiding construction of the theory of causal significance will be *simplicity*. I begin with the question, *What is the least set of features a world must possess to make conventionalism false in that world*? Notice that the concept of causal responsibility comes loaded with *default assumptions* about the character of causal relations. Among these assumptions are the ideas that causal relations are asymmetric, that they exist only forward in time, that they are only local in space, perhaps that they involve events, and that it is a two-place relation.

We can treat these assumptions as default values of parameters on a more basic concept. These parameters are: its arity (how many elements are involved in the

causal connection?); categorical constraints on the relata (do effects and causes need to be events?); symmetry (is the causal connection symmetric or asymmetric?); directionality (if asymmetric, in which direction does the connection go?); and locality (does the connection respect spatiotemporal proximity?). The next step in the analysis investigates whether these parameters need to have any specific values to make conventionalism false in a given world. Taking them one at a time:

The arity of the relation. The arity of a relation refers to the number of things related. The ordinary language idea of causation seems to be of a two-place relation, but conventionalism could clearly be false even if causation were a relation between more than two things. In fact, Evan Fales (1990) has proposed that causation in our world is really a six-place relation between two points in space, two points in time, and two properties.

The categorical constraints on the relata. Hume wrote of causation as a relation between events. Many philosophers, such as Davidson (1967), apparently following Hume, often model it metaphysically as a relation between events. However, Vendler (1962) collected detailed linguistic evidence that in ordinary language it is often a relation between a fact and an event. Also, on the metaphysical level, libertarian philosophers have introduced the notion of agent causation, in which agents are causes. Finally, the tradition of process philosophy, as well as Wesley Salmon's (1984) empiricist view on causation, draft processes as essential elements of causation. One can argue about which proposal best captures causation in our world, but it seems clear that conventionalism could be false regardless of the kind of proposal accepted.

Symmetry. Although our ordinary concept of causation distinguishes between causes and effects, we can imagine a world with symmetric constraints, such as constraints on the simultaneous state determinations of multiple individuals. For example, there could be a world in which a group of tossed coins are constrained to come up in only certain combinations of heads and tails. In these worlds, we imagine the laws of nature ruling out the occurrence of some *combinations* of events, even though each coin, tossed individually, could come up heads or tails. Such a world would not be a conventionalist world because there would be a metaphysical constraint between distinct events. Our world even seems to be such a world, as the quantum constraints on the states of entangled particles rule out some joint instantiations of otherwise possible states. Thus conventionalism could be false even if there were no distinction between cause and effect.

Directionality. Questions about directionality arise only in worlds with asymmetric causal constraints. If asymmetry is not essential to causation, then directionality is obviously not essential to it, either.

Locality. Quantum physics provides reasons for believing that constraints hold between things nonlocally even in *our* world. An objective basis for the existence of such constraints would be enough to falsify conventionalism. In general, any

world in which causal asymmetry is broken in the manner I described earlier could easily violate locality without falling into conventionalism. This phenomenon was first pointed out as a consequence of quantum mechanics in a famous thought experiment proposed by Einstein, Podolsky, and Rosen, later theoretically confirmed by John Bell, empirically confirmed by Alain Aspect, and subsequently reconfirmed by others.

Judging by these considerations, it does not seem as if the parameters on causation need to have specific values to ensure the falsity of conventionalism. Explicit reflections show that our ordinary concept of causation is only one among many possible specifications of a more fundamental and general concept. This more fundamental concept is simply one of real constraint between distinct entities. If a realist view of causation is correct, then the occurrence of an event (for instance) has *significance beyond itself*, a significance that ripples widely through an ontologically interconnected causal mesh, forcing the rest of the world to be, in some sense, compatible with its occurrence. A realist theory of causation will give an account of what causal significance, in this sense, is.

> Definition 9.1: The *causal significance* of a thing is the constraint its existence adds to the space of possible ways the world could be. A successful *theory of causal significance* should lay bare an objective base of facts on which less objective facts about causal responsibility might rest.

Causal significance shows causation to be an operator on a space of possibility. The recognition that a theory of causation can be a theory of causal significance yields a revised decision tree, as depicted in figure 9.2. Causal significance represents the deep structure of causation, and finding a clearer understanding of the deep structure of causation is how a Liberal Naturalist will probe the deep structure of the natural world.

What do I mean by *the deep structure of causation*? By focusing on causal significance, I am suggesting that the causal realist should treat our ordinary idea of causal responsibility as something akin to the surface structure of a grammar. According to one school of thought, the grammar of a specific language is an idiosyncratic development of a more general and universal structure, called the deep structure of language, which is common to it and all other possible human grammars. By analogy, I am proposing that the way we have come to think about causation in our world represents the surface structure of the deeper grammar of causal constraint common to this and all other possible causal worlds. The *deep structure of causation* is the concept of real constraint, conditioned by a variety of parameters whose specific settings represent hypotheses about the structural features that direct the flow of constraint.

9.6 Causal Significance Supercedes Causal Production

A theory of causal significance will have a radically different form than we would expect from a theory of causal responsibility. Theories of causal responsi-

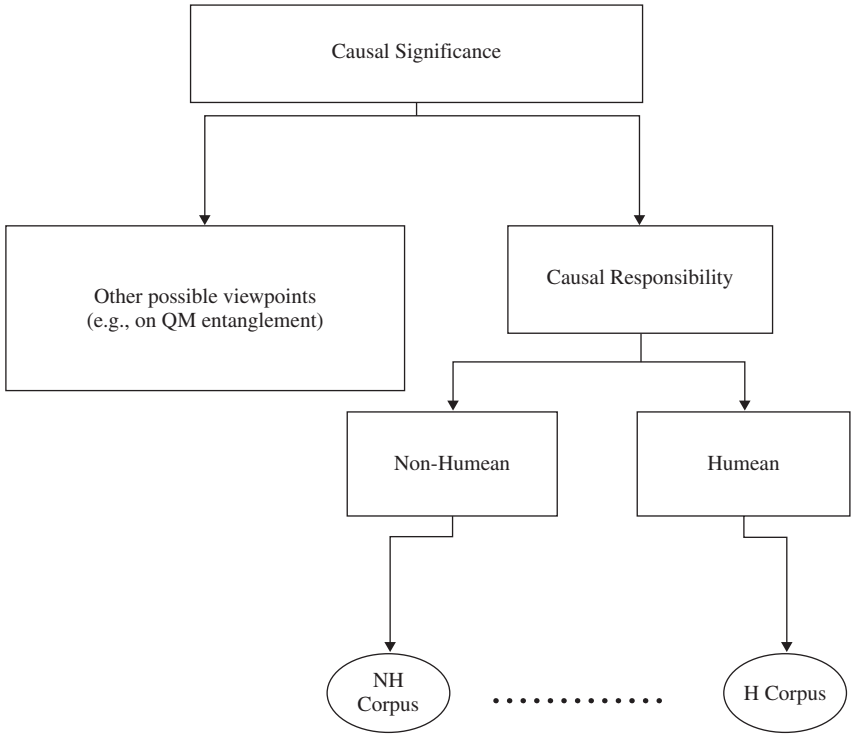


Figure 9.2 Causal responsibility incorporates intentional facts and can be viewed as a refinement of causal significance.

bility invite us not to focus on constraint but on causal powers capable of bringing other things about, on causes producing effects. Therefore, theorists of causal responsibility tend to produce theories of causal production. In a deep sense, theories of causal responsibility start from perplexity that changes occur (why did something happen?), and their driving metaphysical question is the ancient question: *Why is there something rather than nothing?*

In contrast, the core concept behind causal significance is not production. Because production requires one thing to in some sense come "out of" another, production is asymmetric, directed, and naturally limited to local connection. Recall from the previous section that these features of our concept of causation are specific developments of a more general concept of causal significance. If the deep structure of the natural world is a structure of natural constraint, then the logic of constraint leads to a focus on selective inclusion and exclusion rather than production. Conceptualizing the world as the ultimate clique directs questioning to-

ward the discovery of the secret character by which the world denies existence to so many things that could have been.

The humble truth is that, for all we know, existence might be something toward which all things tend. If so, what requires metaphysical explanation might be why some things *aren't* rather than why some things *are*. Perhaps the fact that new things can come into being is part of the noncontingent nature of the world, and perplexity should start at observation of how restrained these facts are in reality. Most possibilities do not occur.

Picture the world as a jewel set in a heaven of transparent possibilities, each flowing along its surface, peering as if at sparkles on ice seen through a window, fingers gently probing for an opening through which it can pour itself. The thought that there could have been nothing becomes strange, and it can seem quite puzzling that there is not much more. Like Robert F. Kennedy, we are not seeing things that are and asking, "Why?" but are dreaming of things that are not and asking, "Why not?"

In a deep sense, the search for a theory of causal significance suggests that the grand metaphysical question all along should have been: *Why is there something rather than everything?* Why doesn't every arbitrary combination of properties occur? This theory of causal significance, the theory underlying the magic of the Canvas of Causation, will be a theory of symmetric and asymmetric *state-constraint* between individuals. It is a theory designed to understand how constraints propagate, so it explains how the actual world comes to be just a sliver of what could have been.

One billiard ball hitting another is a paradigm case of one event causally producing another, and so "billiard ball causation" is not necessarily the best paradigm case of causal significance. A better paradigm case might be two entangled quantum particles. Two entangled particles are similar to two coins that must always be flipped together and that share a special constraint. Although each coin could land either heads up or tails up if it could be flipped separately, making for four possible joint states between them, because the two coins are entangled they share a *constrained joint state* in which each can land heads or tails only if the other one also does. So they could both be heads, or both be tails, but they could not come up one tails and the other heads. In this sense, the state of each has *causal significance* for the other, and their mutual causal significance excludes two possibilities. "Causal significance" names the presence of constraint between them, while not necessarily explaining the state of one by assigning responsibility or temporal precedence to the other.

Causal significance is produced by the set of causally relevant properties an individual possesses. Collectively, these properties constitute an individual's *nomic content*.[1] I analyze nomic content into two fundamentally different but interdependent kinds of properties: the *effective properties* that are responsible for an individual's capacity to constrain the states of other individuals and the *receptive properties* that form a network of connectivity, allowing individuals to place the constraints potential in their effective states.

9.7 Effective Properties

Reflect on how we create physical theories. We are creatures fully embedded within the natural world, and physical theories are our attempts to understand something about the causal order of that world. When we self-consciously consider the position we occupy, the character of the information we gather and hold physical theories accountable for becomes more apparent.

Nature places human beings within an effective loop. We must understand how the world may change us, and we, it. Fortunately, perception provides information to help with this challenge. Through perception we become systematically sensitive to environmental influence, treating some of its effects on us as providing information. Perception selectively processes effects that the environment may have on us, converting those it can into informational fuel that we burn and store in forming our interpretations.

Physicists have strongly tuned the methodology of physics to the effective nature of the world. The genius of the experimentalist is in solving the following challenge: Assuming that the entities we postulate are present, how can we isolate them and identify their states? The basic measuring devices they begin with are those of our biological endowment: eyes, ears, nose, tongue, and touch. The experimental physicist must find ways for perceptible and nonperceptible entities to make a distinctive difference to us via our biological endowment.

For nonperceptible entities, the experimental physicist first finds something else that the ultimate object of investigation can affect; then the experimental setup must magnify this effective difference through a chain. Near the end of this chain is something—perhaps a pointer, a colored flame, a visible vapor, or a computer display—that can affect our senses without the further aid of special instruments. At this last step, the effective natures of our instruments act on our biological endowment, completing the chain. In short, when we measure, we find effects of the hypothesized entities that we can magnify to a level of reality that we can perceive directly. The character of the entire process forces the effective dispositions of things into our theoretical fold because it is always a chain of effects, from hypothesized entities to us, whose explanation we require.

These properties are *effective* because their presence constrains the states that other individuals may also or subsequently have,[2] and experimental science is possible because human beings can arrange and rearrange circumstances so that the total constraint structure changes the state of our biological endowment in systematic ways relative to the property being investigated. With enough information about this systematic variation, we are able to infer the character of the underlying constraints.

In short, the fundamental physics of our universe will be the science that *at least* discloses to us the effective dispositions of the fundamental individuals[3] of our universe, assuming such individuals exist. However, none of this implies that physics will yield a complete account of the world's causal structure. Doubts exist because effective properties require the existence of other kinds of properties.

The three questions that are the focus of this and the next three chapters are: (1) What other aspects of causation exist? (2) How do these different aspects interrelate? and (3) Are these other aspects physical? I argue that causation has two further aspects and that neither is plausibly physical.

9.8 Receptive Properties

This seems to be a *conceptual* truth: A property of an individual may be effective only if some individual is receptive to the property's presence. The two notions, effectiveness and receptivity, are logical complements of one another, so the world cannot realize one without the other. Thinkers in the history of philosophy have often recognized this duality, but usually only briefly and obliquely. For example, in Plato's *Sophist,* the character of the Stranger speaks for the materialists of antiquity, saying:

> I suggest that anything has real being that is so constituted as to possess any sort of power either to affect anything else or to be affected, in however small a degree, by the most insignificant agent, though it be only once. (247e, Hamilton and Cairns, 1961)

Receptivity is something like this *power to be affected* that Plato briefly points to, as does John Locke in chapter 21 of *An Essay Concerning Human Understanding*:

> Power thus considered is two-fold, viz. as able to make, or able to receive, any change. The one may be called active, and the other passive power. Whether matter be not wholly destitute of active power, as its author, God, is truly above all passive power; and whether the intermediate state of created spirits be not that alone which is capable of both active and passive power, may be worth consideration. (Locke, 1690)

This old distinction between active and passive power has fallen to the periphery of modern thinking. Likely, part of the reason is the previously discussed empiricist deflation of causation begun by Hume. Another part of the reason may be the unfortunately oxymoronic name, *passive power*. Despite the empiricist neglect, the idea remains an important part of process philosophy, where process philosophers recognize the logical need for something that does its work (e.g., Griffin 1997).

At times, the conceptual distinctness of receptivity and effectiveness has led us to postulate special kinds of individuals possessing only one of these aspects. For instance, the medieval/Aristotelian conception of God as a purely active force (mentioned by Locke), or unmoved mover, is an isolation of effective properties within a nonreceptive individual. On the other hand, dualist proposals about consciousness are sometimes epiphenomenal. They postulate that phenomenal consciousness is determined by the physical properties of the brain but is nevertheless causally inert. This is the postulation of an individual with properties that are receptive but not effective.

One can intuitively triangulate in on the distinction by considering each case and then identifying the complementary kind of property as what is *missing* in that case. What would an unmoved mover be *missing* so that it, alone among all beings, would be unresponsive? Equivalently, what is it that other beings have that it does not? Answer: It is missing a receptive aspect. What would an epiphenomenal consciousness be *missing* that would make it, alone among all beings, epiphenomenal? Equivalently, what is it that other beings have that it does not? Answer: It is missing an effective aspect.

Because of obvious problems in gaining knowledge about the presence of a purely receptive being, we would not expect any established science to have accepted the existence of one (modulo, controversially, consciousness itself). But has science ever found it intelligible to propose purely effective beings analogous to unmoved movers? Surprisingly, at least one example exists and, maybe, another. The clearest example of a purely effective entity is Newtonian space. Its Euclidean geometry constrained the movement of objects within it, although it was entirely unresponsive to its occupants. From the perspective we are now discussing, the causal difference between Newtonian space and Einsteinian space is twofold. First, the introduction of a different geometry represents a change in its effective nature. Second, Einstein introduced responsiveness to the distribution of mass within it. This second change is an entirely different kind of addition, ontologically, and the more revolutionary. Einstein added *receptivity* to space.

Although Einstein robbed Newtonian mechanics of its only unmoved mover, he may ironically have introduced another kind of his own: singularities. As entities with infinite density, singularities seem to have great effect on the rest of the universe. For instance, they create black holes. On the other hand, it is not clear that anything can, even in principle, affect them in return. Singularities may lack receptivity.

Collectively, these examples show the conceptual and empirical distinctness of effectiveness and receptivity. This distinctness marks an important point: They are not identical aspects of causation. These two aspects of the causal process do different jobs, and they need distinct accounts. A proper account will detail how each aspect helps to ground the very possibility of causal activity. Importantly, each aspect presupposes the possibility of the other's existence, so the conceptual relation between these two aspects of causation, the effective and receptive, has a circular structure. They are thus interdependent and equally fundamental aspects of the causal nexus.

I will revisit the case for receptivity in chapter 13, summarizing both these philosophical reasons for accepting its existence and further empirical reasons given in the next few chapters. For now, we know that (1) we should interpret physical theory in a causal realist way; (2) the ideal physics will include *all* the effective dispositions of our world's fundamental individuals, and (3) the effective and receptive aspects of causation are conceptually and empirically distinct.

Points (1)–(3) have already been established. For a moment, I assume something that I will argue for later, that (4) physics exhibits *only* the chain of regu-

larity between instantiations of the effective properties. If all of (1)– (4) are true, it follows that causation in our world has at least two equally fundamental aspects, and that one of them, receptivity, is left out of physical theory. Receptivity is an explanatory luxury for physical science, but it is nevertheless metaphysically relevant to the causal structure and evolution of the world.

If premise (4) is true, the overall ontological picture becomes very interesting. Receptive properties are necessarily related to the physical in that the physical properties are only effective properties, and something's being effective presupposes something's being receptive (and vice versa). In a world that realizes effectiveness, we have a necessary coinstantiation of logically distinct essences. Nevertheless, the logical connection between these aspects is not one of supervenience (because it is mutual), and the necessity connecting them is not merely nomic (because it is not logically contingent). It is a natural dualism of necessarily connected dualities, but not one that involves a merely nomic, external connection. We are on the cusp of a significant metaphysical proposal for the nature of causation that takes us beyond physicalism.[4]

9.9 Receptivity as a Connection

If receptivity itself provided a connection between individuals, it would support a metaphysically far richer theory than a simple sponge metaphor in which receptivity is just another kind of monadic (i.e., one-place) property. Figure 9.3 visually contrasts the two alternative pictures. In this section I develop a connectivity view of receptivity.

I have several reasons for eschewing the monadic alternative and preferring to model receptivity as a connection. One reason is that, if one adopts the monadic view of receptivity depicted at the top of figure 9.3, the problem of "activating" an individual's receptivity relative to the effective states of other individuals remains. An individual cannot just be receptive *simpliciter*: It must be receptive *to* the effective state of some other individual(s). To complete the account, we would have to specify some conditions for selectively determining which individuals a given individual will be receptive to. This further condition, whatever it might be, is a complication to the model that does not arise if one begins by modeling receptivity as a connection.

Aside from the inelegance this extra step introduces, it also tends to limit the account in unnecessary ways. For instance, the tempting further condition is the classical assumption of spatial or temporal contiguity. This classical move rules out nonlocal causal connection by definition, which seems undesirable. It also brings spacetime into the picture in a fundamental role, precluding the otherwise attractive possibility of reducing it to more fundamental facts about causal connection.

A second reason for preferring the connection view is the very elegant modeling of levels of nature it allows, at least with respect to the emergence of higher level individuals incorporating lower level individuals. What the connection view

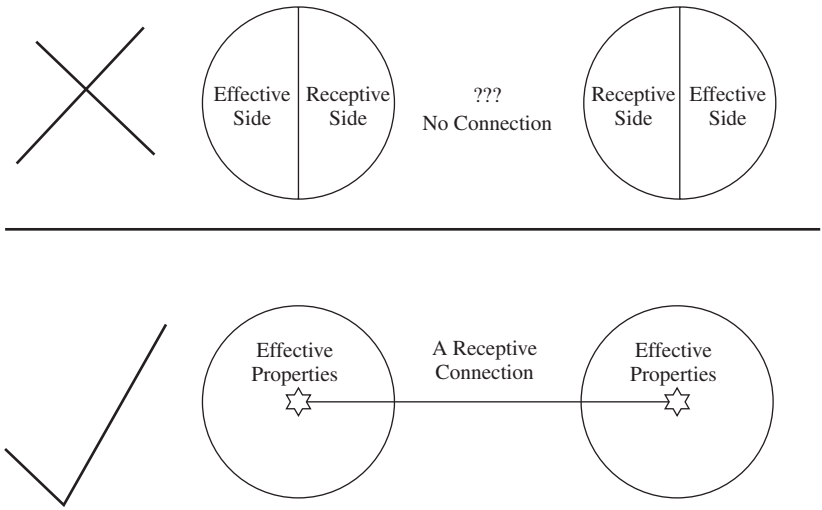


Figure 9.3 The Theory of Causal Significance will suppose that receptivity is a special kind of connective property different in some crucial ways from a traditionally conceived monadic property.

offers is an opportunity to specify the conditions of a substantial internal unity—a shared receptivity by multiple individuals—that may ground a notion of natural individual and natural individuation. Receptive connections, and thus natural individuals, could exist at any level of nature, and so there would be no special pride of place given to microphysical individuals. This feature dovetails well with modern science, as seen earlier in Michael Lockwood's observation that:

> In quantum mechanics there is a sense in which all observables, and in particular observables corresponding to every level of structure, are to be regarded as equal in the sight of God, as are different frames of reference, relativistically conceived. As I intimated earlier, quantum mechanics seems to be telling us that it is a classical prejudice to suppose that the world is not intrinsically structured at anything but the level of elementary particles, and their actions and interactions.

A final reason for preferring the connection view is that it produces a causal mesh with a topological structure of its own. Connections could be either symmetric or asymmetric, and the receptive face of causation would have the form of a directed graph providing a kind of scaffolding off which the rest of nature could hang. Both a theory of causation and a theory of consciousness must eventually grapple with issues involving the nature of space and its relation to time. The topology provided by connectivity gives hope for grounding a reduction of the facts about space and time, potentially increasing the explanatory power of the theory. The spatial assumption of such a reduction would be that there is *a causality condition on locality, not a locality condition on causality*. With respect

to time, temporal succession and precedence would correspond in a structured way to asymmetries in relations of causal constraint.

9.10 The Theory of the Causal Nexus

The determination problem. Assume we accept that causation is not fundamentally about causal production. What does it mean to say that, instead, causation is about constraint on a space of possibilities? What problem does the causal nexus resolve for nature?

I will frame the alternative to causal production through reference to the *determination problem*. The determination problem arises from the fact that the world's individuals each have many potential states. To be actualized, an individual must take on one, and only one, of these potential states. That is, it must become *determinate*. One can imagine the world's basic properties, say, mass, charge, and velocity, as mere potentials with many possible determinate values. The determination problem is to create a determinate world from these indeterminate potentials. Causation solves the determination problem.[5]

The determination problem does not necessarily have to be solved by a process through time. Introducing levels of constraint is another possibility. If one thinks of causal connection as an operator on a space of possibility, one can imagine that it is applied in the basic instance to basic determinable properties at a moment in time. If this first-level application does not yield a completely determinate state for the world, a second-level operator can be applied to the results of the first-level operator, and a third-level operator to the results of the second-level operator, and so on in layers until the total set of causal connections in the world produce determinacy. As I explain in detail here and in the next chapter, these successive applications can occur at successive levels of organization as easily as they might occur at successive moments of time.

In the classical scientific and commonsense picture, causation solves the determination problem almost at once: The interactions between basic particles and forces constrain them to have determinate states. This classical viewpoint is a two-level solution in the sense that the constraints on particle natures count as one level of constraint and that their interactions through forces count as a second level of constraint. After the second level of constraint the lowest level entities are determinate and there is no more need for causation: The determinateness of things at higher levels is a direct consequence of the determinateness of things at the lower levels.

Although intuitive, almost dangerously so, this classical conception is not an a priori truth. Nature might solve the determination problem in one or two levels, as the classical conception presumes, or nature might have to add further layers of causal connection before the determination problem is resolved. If that were true, it would be counter to classical views of the world, but not unsupported by evidence or wholly surprising. We actually have some a posteriori reasons from quantum mechanics to believe that the classical presumption is false and that the

lowest levels of constraint leave the states of the lowest level individuals in the world indeterminate. The question of whether this quantum evidence is what it seems to be, or whether the classical view will win out in the end, is open to discussion and further evidence. If the classical conception were false in a world, it would imply that layers of fundamental causal relation above the lowest levels are needed to make that world's individuals fully determinate.

Overview. The theory I develop below is a theory of the causal nexus. It allows us to model classical and nonclassical solutions to the determination problem, and it is explicitly agnostic about how many levels of causal connection are in the actual world. The theory's purpose is to provide a framework in which one could model many proposed answers. Thus perhaps causation solves the determination problem by taking one, two, or two hundred steps up the ladder of nature. From the perspective of the theory here, all answers are equally acceptable. Its concern is to allow the questions to be posed by representing more general truths about what causation is and how it works.

My first step will be to give a very high-level gloss on the overall shape of the theory. I will do this by introducing a few basic definitions and by propping up an example of how, on the view of causation to be developed, the determination problem might be resolved for neural states at a middle level of nature. In a causal realist's world, there will *at least* be:

Definition 9.2: *A causal nexus* (pluralized as causal nexii)–A receptive connection binding two or more effective individuals.
Definition 9.3: *Effective properties*—Properties that contribute to constraints on the determinate states of a causal nexus.
Definition 9.4: *Receptive properties*—Connective properties enabling individuals to become members of causal nexii and to be sensitive to constraints on the state of nexii where they are members.
Definition 9.5: *Causal laws*—Laws describing restrictions on the composition of the causal nexus; that is, laws describing the compatibility, incompatibility, and requirement relationships between effective properties within a nexus.

These four commitments form the skeleton for a theory of nomic content and, therefore, of causal significance. A theory of causation will come from more fully articulating and tightening these skeletal ideas.

For a first pass at tightening these ideas, I am going to gloss a hypothetical causal life and causal context for an arbitrary neural cluster. The purpose of this first example is to gradually introduce the way of thinking suggested by the determination problem and embodied in the theory. The example illustrates some general principles and asserts some of the key concepts without introducing too much detail. The detail and explanation will come later.

How can we become accustomed to thinking in terms of the determination problem? Imagine a neural cluster *NC* that is one of sixteen such clusters NC_1 to NC_{16} densely interconnected in the brain. How might we understand their causal

relations if the determination problem has not already been resolved at a lower level of nature? Before we can say much to answer this question, we first need a clearer way of thinking about what it asks, so before describing the relations between these clusters I define two new concepts.

The first concept is that of something having a state "considered independently" of its environment. The state of an individual *I,* considered independently of its environment, is the state it could be said to have if one took account only of the causal relations internal to it, that is, the causal relations possessed by its own constituents. In the context of the determination problem, this is a way of asking whether the causal constraints held solely by its constituents are strong enough to produce a determinate state for *I*. So, for a given individual *I*, even if *I* is in a determinate state given the whole causal situation in the world, there is a question to ask about whether it is determinate "considered independently" of its causal relations to its environment. This question can have either a yes or a no answer. We can therefore define:

> Definition 9.6: *I* is in a determinate state when *considered independently* if, and only if, the causal relations belonging to the constituents of *I* entail that *I* is in a determinate state.

There are two conditions under which *I* would be determinate when considered independently. These two conditions are (1) its constituents are each already determinate considered independently or (2) the existence of *I* itself adds some causal relation among its constituents that makes them determinate. In all other cases, *I* would be *indeterminate* when considered independently.

Closely related to the concept of an individual *I* being considered independently is the concept of the states that are *independently possible* for *I*.

> Definition 9.7: A state *S* is *independently possible* for *I* if, and only if, *S* is a state left open for *I* when *I* is considered independently.

If an individual *I* is determinate when considered independently, then there will be only one state *S* that is independently possible for *I*. However, if *I* is indeterminate when considered independently, then there will be more than one state *S* that is independently possible for *I*.

Given these definitions, suppose that *NC* and the other neural clusters are each in indeterminate states when considered independently. It follows that:

1. There are many independently possible states for each of them.
2. Considered independently, the number of possible joint states of the neural clusters is the Cartesian product $NC_1 \times NC_2 \times NC_3 \times \ldots \times NC_{16}$ of their individual independently possible states.

This is what it means to say that the determination problem has not been resolved for NC_1 through NC_{16}. In fact, if even one of the clusters were in an indeterminate state when considered independently then we could not say the determination problem was resolved for the group of clusters, as they are densely

interconnected and we can assume their joint state is critical for other systems. If even one cluster were indeterminate when considered independently, then the joint state of all the clusters potentially would be indeterminate with consequences for any further systems whose behavior might depend on their joint state.

With this aspect of the determination problem understood more clearly, we can ask again, How might we understand their causal relations if the determination problem has not been already resolved at a lower level of nature? The purpose of a causal relation is to help resolve the determination problem so it seems that here there is work to do for a basic causal relation. Receptivity will stand in as this basic causal relation.

Please recall from the previous section that I am going to treat receptivity as a connection: Each instance of receptivity can be shared in common by multiple individuals. With this in mind, assume that NC_1 through NC_{16} share a common receptivity. Here, please consider their common receptivity to be a novel ontological factor not derivable from lower level conditions. Through sharing it they are bound together within a single causal nexus.

The theory attaches two kinds of significance to the sharing of this common receptivity between *NC* and the other neural clusters. First, each cluster is an individual in the nexus and there are conditions, described by causal laws, for cohabitation of a single causal nexus by multiple individuals. The existence of causal laws means that the states available to each neural cluster within the nexus are directly constrained by whatever effective properties the others possess. Second, their shared receptivity establishes the potential for them each to be part of a common receptive field with the others. Within this common receptive field their joint states could be constrained as a whole by interaction with external influences. The facts of the situation are depicted in figures 9.4 through 9.7.

Figure 9.4 simply depicts *NC* as a neural cluster.

Figure 9.5 represents five independently possible states for the neural cluster *NC,* each state represented by a different shading, depicting the fact that *NC's* internal causal relations do not constrain it to a unique state. When a situation like this is true of an individual like *NC*, I say that the individual is indeterminate when considered independently.

Figure 9.6 represents a shared receptive connection between *NC* and other neural clusters. This connection represents a causal nexus that *NC* has entered into with the other clusters. These other clusters represent *NC's* receptive field. The other clusters sharing this receptivity provide an immediate environment for *NC* at its own level of organization, and *NC's* environment adds constraints to its state over and above those it has when considered independently of its environment. By taking on environmental constraints, *NC* may find that some of its independently possible states are no longer open to it.

Figure 9.7 shows the whole group subject to a common receptive field at a higher level of organization. Just as *NC* has a receptive field of its own, which consists of the other fifteen clusters to which it is connected, it is also part of a

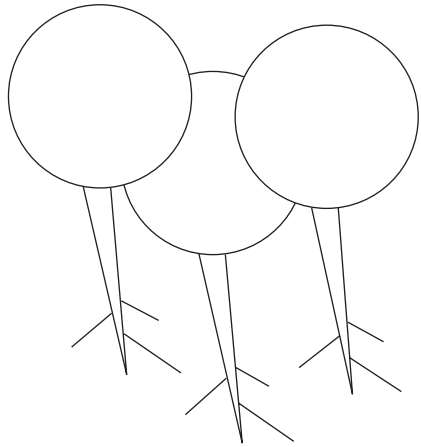


Figure 9.4 A neural cluster *NC*

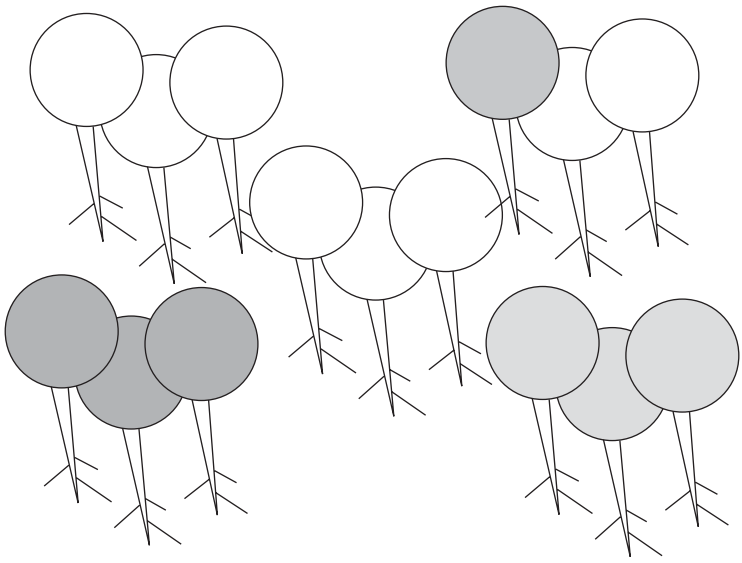


Figure 9.5 Indeterminacy in *NC*: Each shading represents a different possible state for *NC* when considered independently of its environment.

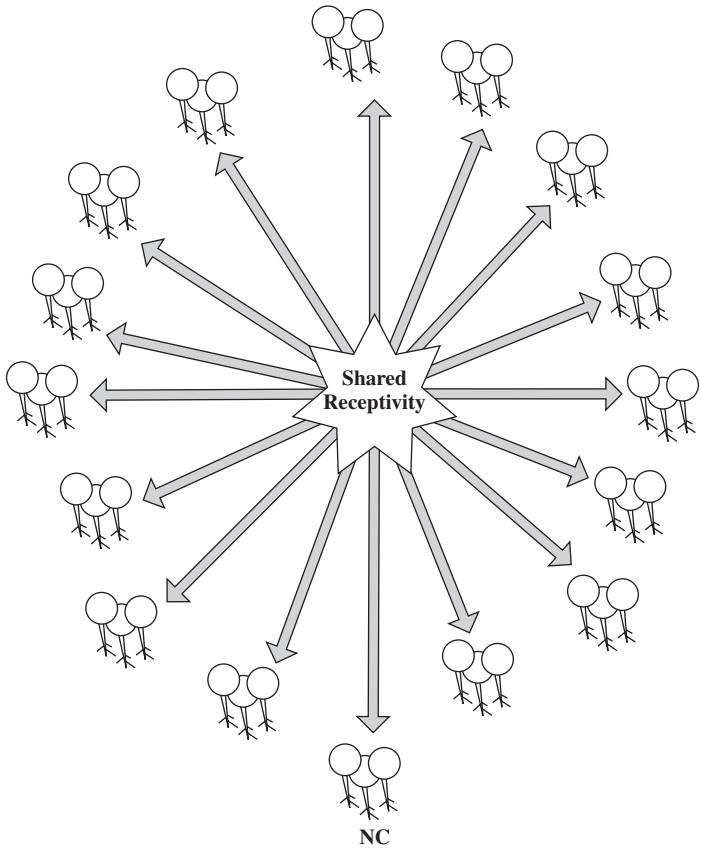


Figure 9.6 Direct causal influence between the different neural clusters arises through a shared receptive connection. The influence of the other clusters presents further constraints on the possible states open for *NC*.

collective supercluster emerging from the shared receptivity of the sixteen lower level clusters. The supercluster also may have its own receptive field, enabling further environmental constraint on its state.

It is a basic tenet of this view that, as a consequence of the common receptivity shared between *NC* and the other members of the nexus, there is a common constraint structure that reduces the space of their possible joint states. Furthermore, in the context of its shared receptivity with the other clusters, *NC* is no longer being considered independently, and we can suppose that the elimination of some possible joint states for the network of clusters results in the elimination of some of *NC's* independently possible states.

For the sake of the example, assume that only one of *NC's* independently possible states remains in the set of permissible joint states. As a result, *NC* becomes

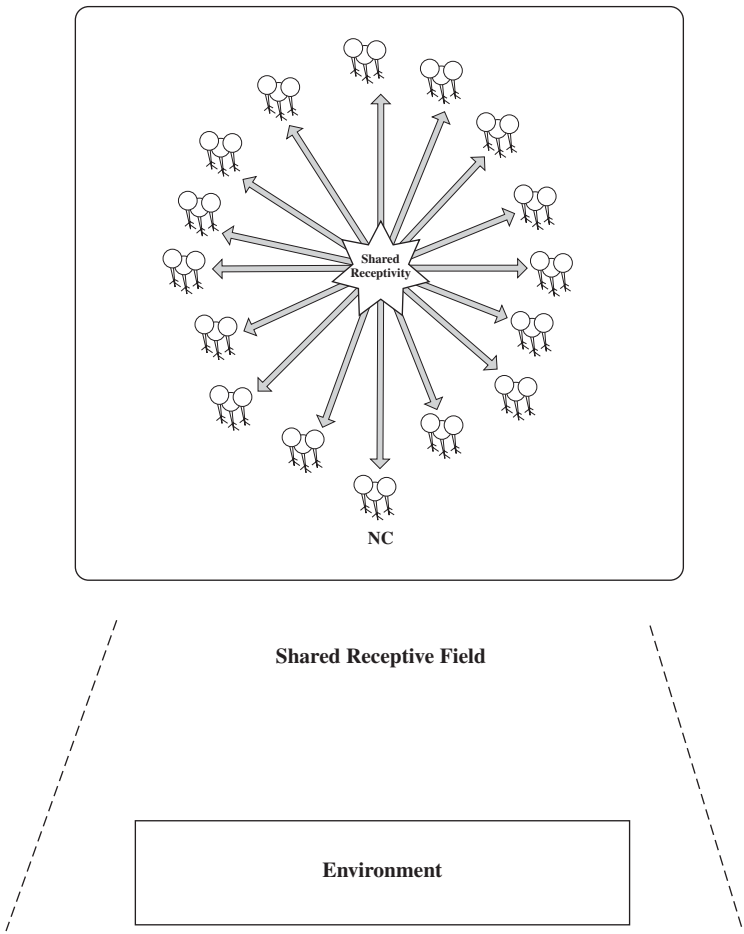


Figure 9.7 The existence of a shared receptivity establishes a common receptive field for the group of neural clusters.

determinate, and the determinate state that *NC* finally manifests is the result of influences active on its entire situation within the nexus: The individuals in the shared receptive connection add constraints to the possibilities for *NC* relative to what they might be otherwise.

9.11 The Deep Structure the Causal Nexus

Binding. Having glossed the high-level story, I can begin to examine the low-level detail. Note here that the particular concept of "individual" being used by the skeletal commitments needs further definition. Because simplicity is the pre-eminent virtue guiding construction of this fundamental theory, I keep strongly to

parsimony constraints. My most primitive individuals are just the most primitive effective properties (e.g., Mass, Charge, and Spin) and receptive properties. In the theory these property instances are called level-zero individuals as illustrated in figure 9.8.

I develop a view whereby receptive connections are special *properties* whose instances can *bind* to more than one individual at a time. The individuals a receptive property binds, together with the receptivity, create a new individual. We can consider this new individual to be a level-one individual constituted by the binding of the level-zero individuals. This new individual is the one to whom the receptivity *belongs* in the more conventional sense of a property belonging to an individual. The level-zero individuals in general belong to the level-one individual constituted by their binding. Level-one individuals might be things such as the fundamental particles. Figure 9.9 illustrates the creation of this kind of complex level-one individual from the binding of the simple level-zero individuals.

Formally, if a two-place receptive connection *RP* binds to two primitive effective properties EP_1 and EP_2, together they form a higher level individual (e.g., a fundamental particle) that has as a property the receptivity *RP* and an effective state consisting of EP_1 and EP_2. The principle generalizes for receptive connections of more than two places and, with respect to receptivity, for individuals at higher levels than level-zero (discussed later).[6] For example, in applying this principle to the previous discussion of *NC*, we would say that their common receptivity *binds* each of *NC* and its fellow neural clusters. Thus the sixteen clusters together come to constitute an individual that has the receptivity as a property, leading to the possible existence of a receptive field for the new individual.

However, I tread carefully, because any understanding of the effective and receptive properties must respect the special categorical interdependence between them. To represent this interdependence, I propose thinking of the properties themselves as having incomplete natures and needing to bind with individuals possessing the complementary kind of property to complete. This *binding rela-*

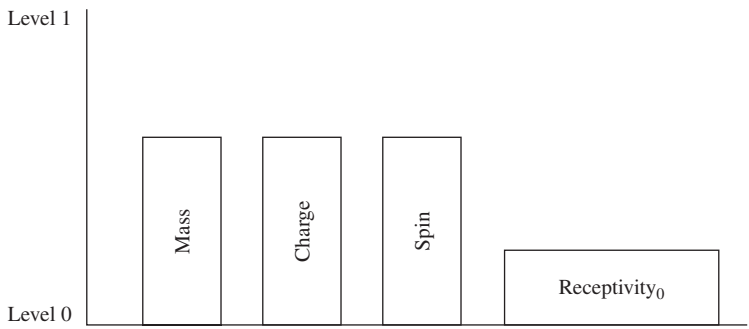


Figure 9.8 Four basic properties, three basic effective properties and an instance of receptivity, existing as level-zero individuals.

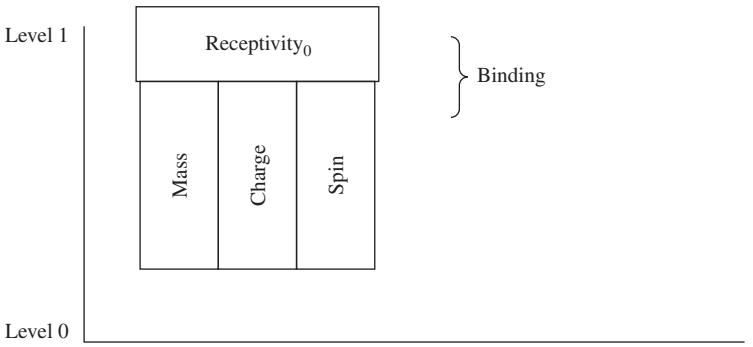


Figure 9.9 The level-zero effective individuals may bind to the level-zero receptivity, creating a level-one individual such as a basic particle to which they all belong.

tion must be a unique kind of *internal* relation between the effective and receptive properties that allows for a kind of metaphysical completion of their essences. When incomplete natures bind to one another, the binding achieves three things:

- First, portions of bound properties become *part* of the incomplete natures they bind to, making those natures more complete.
- Second, a collection of bound natures containing more than one effective individual becomes a *causal nexus*.
- Third, binding supports a kind of transitivity, and so it provides the mechanism of causation by enabling the penetration by which distinct effective natures can *allow*, *include*, or *exclude* one another.

The thesis that completion through binding enables a kind of transitivity is important and it is illustrated in figure 9.10. In figure 9.10 the three effective properties Mass, Charge and Spin are shown as taken up, through binding, into the completion of the receptivity R_0, which in turn is shown as part of the completions of the three effective properties. Through R_0 each of the effective properties, or some part of their individual determinable natures, becomes part of the completion of the other two effective properties.

To illustrate the importance of transitivity, imagine that through binding some part of an effective nature E_1 becomes part of the completion of a receptive nature R. For the example, assume that E_1 is already complete so that R does not become part of its completion. As a connective property, R becomes part of the completion of a second effective nature E_2, and, because E_1 is part of the completion of R, by transitivity E_1 becomes part of the completion of E_2.[7] R then constitutes an asymmetric connection between E_1 and E_2. It is at this point that the internal relations between effective properties become relevant. One effective property cannot form part of the completion for another effective property unless

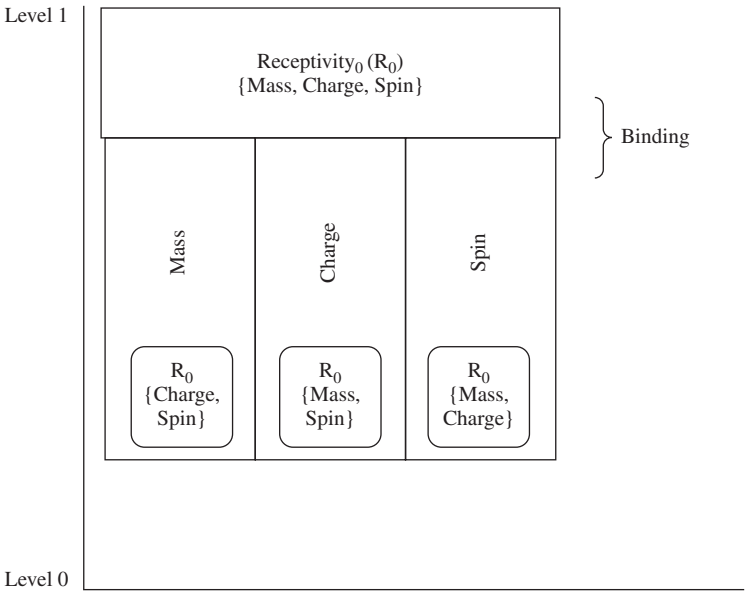


Figure 9.10 An illustration of how binding achieves completion by having the distinct receptive and effective essences penetrate one another, and also how completion supports a kind of transitivity in which distinct effective properties can become parts of one another's completions.

the nexus satisfies the conditions of their internal relations. For example, if through a shared receptivity *R* an effective property E_1 becomes part of the completion of a second effective property E_2, and a potential value for E_2 is incompatible with the value of E_1, then that potential value will be eliminated from E_2*'s* determinable nature. This is a case in which I say that E_1 *conditions* E_2.

The introduction above is enough to suggest the importance of transitivity, and there will be more detail about conditioning later. Let us now continue the introduction by focusing on the fundamentally important ideas of incomplete natures and their completion. Recall that a determinable is a general property such as *redness,* which can have a variety of possible shades, called its *determinates*. Similarly, *shape* is a determinable property that can have a variety of determinates: triangularity, squareness, rectangularity, hexagonality, and so forth. Physical properties such as charge are also determinables with determinate values such as positive and negative. Incomplete *effective* natures and their *completion* follow the model provided by the traditional concept of a *determinable* becoming more *determinate*. Effective properties are determinables, and their completion is a process of their becoming more determinate.

One can think of an incomplete effective property as a determinable of a sort. It is an abstract entity that contains a propensity within it to become one of its

determinates. Depending on the character of the determinable, these determinates are shapes or forms or qualities or quantities that the abstract nature may take on completion.

As for the receptive properties, I propose thinking of incomplete receptive properties as neutral essences with a kind of inherent openness representable as a set of "slots." These "slots" accept effective individuals to which the receptive property binds. To give some imagery to it, think of effective individuals as cans of Coke and receptivity as the thin transparent plastic that binds cans of Coke into six-packs. The loops in the plastic that bind the Coke cans are like the slots in the receptivity to which natural individuals bind. The idea of a "slot," then, is a metaphorical way to represent a receptive connection's *carrying capacity*. In this book, I always represent a receptive connection as having a discrete and finite number of slots, although I believe the theory could be extended to instances of receptivity with a nondenumerable capacity.

Whereas the plastic binding of a six-pack merely curls around Coke cans, a receptive connection binds individuals in a much deeper and more penetrating way. Binding is an internal, metaphysical relation between abstract essences (i.e., the otherwise incomplete effective and receptive natures). When effective and receptive natures bind, I say that the corresponding receptive slots become *saturated* ("saturation" is analogous to the plastic loop being "filled") by the effective determinable and do not merely hook around it externally. The saturation constitutes a merging of the two natures into a more complete nature.

In binding, each incomplete member becomes more complete by taking up some part of the nature of the thing to which it binds.[8] So some part of the effective determinable becomes part of the receptive openness, and some part of the receptive openness becomes part of the effective determinable. If two or more effective natures bind to the same receptivity, then I say that they share a common receptivity, and the new entity forms a causal nexus.

> Definition 9.2 (expanded): *Causal nexus*—Two or more nonneutral determinable individuals (i.e., effective individuals) sharing a common neutral essence (i.e., a common receptivity). A causal nexus must have exactly one receptive connection binding more than one effective individual.

For reasons I will discuss in detail in the next chapter, the unbound incomplete natures are abstracts, and so the causal nexus is the basic kind of individual inherent in nature. If this is correct, it follows that instances of pure effective or receptive properties do not exist in nature, and so there are no pure level zero individuals realized in the natural world. They are only metaphysical abstracts. Instead, nature contains effective/receptive complexes.

In some (shallow) respects, the relationship between effective and receptive properties is like the relationship between the front and back of a wall (assuming for the sake of analogy that "front" and "back" name absolutes). The two faces of a wall are distinct, just as receptivity is distinct from effectiveness. Yet a wall cannot exist without having both a front and a back, just as a natural individual

cannot exist without both effective properties and receptiveness. Also, in a generic sense, the front and the back of a wall are necessarily connected: It is impossible that the front of a wall should exist without a back of the wall existing, and vice versa. The relationship is one of mutual necessity and is neither supervenience nor identity, just as the existence of effective properties and receptivity mutually necessitate one another, although their relationship is neither one of supervenience nor identity. Also, it is natural to think of the front and back of the wall as being two aspects of the wall, just as it is natural to think of effectiveness and receptiveness as two aspects of a natural individual. Yet underlying the two aspects of the wall are two surfaces possessed by the two faces of the wall, one face which has the property of being its front and the other face which has the property of being its back. Similarly, the effective properties and receptive properties are distinct properties underlying the different aspects of an individual's nomic content.

Notation. My notation models these effective/receptive complexes. An incomplete receptive/effective complex is a nature denoted by expressions such as *EP*(_,_,....,_), where *EP* by itself would denote an effective property (or an individual with effective properties); (_,_,….,_) by itself would denote an open receptivity; and *EP*(_,_,....,_) denotes the effective/receptive complex created by *EP* binding to the receptivity. Returning to our Coke metaphor, *EP* is like a can of Coke, and the underscores in between the parentheses represent unfilled loops in the plastic binding used to hold the six-pack together.

Because of their internal relations of compatibility, incompatibility, and inclusion, effective individuals have a feature that is not present in the image of the six-pack of Coke: The bound effective individuals each contribute to a set of state constraints on the nexus (i.e., on the six-pack). These state constraints determine what determinate features the members may have. Imagine that each can of Coke is initially a blank tin with many different designs potential within it and that what design finally graces the can depends on which other cans are bound into the six pack. As cans are bound with one another, definite features begin to appear: the Coca-Cola logo begins getting more and more distinct, the ingredients list begins to fill out, and red appears. The appearances of the can's design features are like the determination of an individual's effective properties.

The state of the nexus is the joint state of its members, so the set of state constraints to which each effective individual contributes is a set of constraints on the joint states of the members of the nexus. Depending on the nature of the shared receptive connection, this constraint placement might be *asymmetric* or *symmetric*. If it is an asymmetric connection, then the constraints are structured so that one or more individuals constrain the states of one or more others but do not have their states constrained in return. If the connection is symmetric, then the constraints on the state of the nexus may affect every individual bound to the connection.

I represent the *asymmetric* binding of an effective property EP_2 to an effec-

tive/receptive complex such as $EP_1(\_)$ as $EP_2 \Rightarrow EP_1$, signifying that EP_2 has saturated the open slot in the effective/receptive complex denoted by $EP_1(\_)$ and is now constraining EP_1. If EP_1 and EP_2 share a *symmetric* receptivity that creates a symmetric constraint between them, I abbreviate this as $[EP_1.EP_2]$ to reflect the reciprocal relation between the effective properties in sharing the receptivity. Complexes of more than two effective properties all sharing a common symmetric receptivity would be represented by notations such as $[EP_1.EP_2.EP_3]$, $[EP_1.EP_2.EP_3.EP_4]$, and so forth.

Primitive natural individuals. My proposal for understanding primitive natural individuals is that the primitive level-zero natural individuals bind together to compose the most basic effective/receptive complexes such as $EP_2(\_)$, $EP_2 \Rightarrow EP_1$, and $[EP_1.EP_2]$: i.e., The pure effective determinables (e.g., EP_2) and the pure open receptivities (e.g., (_, _)) bind to become the basic effective/receptive complexes. As stated earlier, these pure level-zero individuals are abstracts, and are never found in a pure state in nature. It is as if the government prohibited the Coca-Cola Company by law from selling single cans of Coke or distributing completely unfilled plastic binders. On this analogy, level-zero individuals are like loose singles of Coke and empty plastic binders that can never make it out of the warehouse and into the marketplace. Instead, it is essential that level-zero individuals be bound to one another in complexes where the receptivity is saturated and the determinable can be made more determinate.

These complexes are *pure property complexes* constituted by (1) one or more effective determinables and (2) a receptive openness binding them directly to itself and indirectly to one another. Furthermore, when a level-zero instance of receptivity has all its slots saturated, the resulting causal nexus such as $[EP_1.EP_2]$ constitutes the creation of a *level-one individual* made from the level-zero individuals by the special binding relation holding between their natures.

Definition 9.8: A receptive connection is *complete* if, and only if, it does not contain an open slot.

Definition 9.9: *Level-one individual*—A completed receptive connection consisting of a level-zero receptivity binding level-zero effective properties.

Causal laws. The resulting ontology is an event ontology in which the actualization of an individual is the fundamental natural event and in which individuals may be internally linked into processes. Individuals themselves are pure property complexes (i.e., there are no enduring substances). Descriptions of the restrictions on the composition of a causal nexus are *causal laws* (i.e., laws describing the possibility of immediate causal connection between individuals). Causal laws, then, are *laws of completion for a causal nexus*. I introduce "causal laws" as a technical term here. Causal laws are not descriptions of regularities in the instantiation of properties through time, which are what we traditionally have called the laws of nature or laws of physics.

I illustrate causal laws by recalling the imaginary example of the two coins

that must be flipped together and that share a joint state. Recall that the constraint on their joint state is that they both have to land heads up or both tails up; one cannot land heads up and the other tails up. Using the apparatus being introduced here, a coin's potential to land heads up or tails up is analogous to two determinate states of a determinable property that the coins may have. Call this determinable property its *landing property*. The constraint on the joint state of the coins would be associated with (1) the existence of a shared symmetric receptivity binding the two coins within a nexus and (2) a causal law describing how their individual landing properties are mutually compatible or incompatible. In this example, the causal law describes the conditions under which different instances of the landing property can coexist within the nexus. A causal law sufficient to describe the behavior would be: *A heads-up value of the landing property is compatible only with another heads-up value*.

Effective natures sharing a receptive connection contribute to global constraints on the state of the nexus. The contributions of different members of the nexus may be seen as either *completely* or only *partially* constraining other members of the nexus. The example of the two coins illustrates at least potential complete constraint in the sense that any determinate value either coin takes for its landing property completely determines the value the other coin must take. It is also a case in which there are two independently possible states for the linked coins together. Therefore, the definite state they take on must be determined by wider conditions to which they individually or collectively become bound.

Partial constraint is more relaxed than complete constraint. If members partially constrain one another, their copresence within the nexus means that particular determinate values they may take on may exclude some, but not all, of the latent potentialities within the determinable natures of other members. To illustrate further, imagine that we had two six-faced dice similar to the two coins in that they are bound to a common symmetric receptivity. This means that there is a constraint on their joint state. Imagine also that the causal law describing the restrictions on the landing properties of these dice is that one die landing with an even number on its face is compatible only with the other die also landing with an even number on its face. In contrast to the compatibility relations between values of the landing property on the coins, the value of the landing property of one die would only partially constrain the value of the landing property of the other die. So a die landing with a six on its face leaves three possibilities for the other die: two, four, or six. Even given the value of one die, the value on the other die is left indeterminate.

What if wider conditions binding one of the coins or one of the dice were to fix the value of that coin or die, say forcing a coin to land heads up or a die to land with the number two face up? The coin whose landing property was fixed by other circumstances to be heads would fully constrain the other coin to be heads also (If the coins are taken to be analogous to entangled particles, we can imagine this as a circumstance in which one of the particles is measured.). The die whose landing property was fixed by other circumstances to be two would

partially constrain the other die, leaving only two, four, and six as possible values for its landing property. Whenever one property or individual in this way fully or partially constrains the state of another, I say it *conditions* that other property or individual, where this conditioning corresponds to making the determinable more determinate by narrowing the set of potentialities within its nature. Notice the role receptivity plays by connecting natures of effective determinables so that they may condition one another:

> *Receptivity itself acts as the causal connection. Nature needs no other ontological grounding for the causal connection.*

Higher-level individuals. Because the causal connections between individuals at a single level might only partially condition one another, there might be a hierarchy of natural individuals. The possibility for further stratification would exist whenever the effective state of the level-one individual was still indeterminate in some respects. In general, partial determination would occur if a determinable property *EP* held multiple determinate potentials in its nature, for example, 0, 1, 2, and 3, and if it bound with a receptivity whose other bindings exclude only some of those values, for example, 0 and 1. In such a case, the level-one individual would still have a determinable state containing values such as 2 and 3 as possibilities for *EP*.

This is like the example of the two dice. A roll of the dice does not, by itself, contain enough constraint to determine the joint state of the dice or even the individual states of either of the dice. However, it is possible that the individual that is the two dice together could belong to an environment of other natural individuals whose presence adds further constraints and succeeds in determining the joint state of the dice.

The relevant indeterminacies correspond to remaining incompletenesses in the effective nature of the level-one individual. In such a case, the individual is still an abstract in some respects and, as such, is still a complex of potential rather than a fully concrete determinate. The determination problem is not yet resolved for that individual, and causation has more work to do. To become fully determinate, the level-one individual would need to bind within a causal nexus with other level-one individuals to form a *level-two* individual analogously to the way that the level-zero effective properties form level-one individuals.

Figures 9.11 and 9.12 illustrate the creation of a level two individual in this manner. Figure 9.11 shows two level-one individuals, again visualized as some sort of elementary particles, at least one of which we can assume is indeterminate when considered independently. Figure 9.12 shows a level-one receptivity binding them together into a level two individual, with that receptivity belonging to the newly constituted level two individual. The earlier remarks regarding transitivity continue to apply, and so we can assume this new nexus has constraints of its own that help resolve the determination problem.

The general idea here suggests an intuitively plausible principle linking completeness with determinateness:

Determination indicates completion. When a determinable nature is complete, it is fully determinate.

The principle that determination indicates completion suggests two further definitions:

Definition 9.10: An effective property is *complete* if, and only if, it is in a fully determinate state.
Definition 9.11: A compound individual is *complete* if, and only if, all of its member individuals are complete.

The principle that determination indicates completion also suggests a basic causal postulate:

The principle of maximal completion. Individuals seek completeness.[9]

The principle of maximal completion names a tendency without implying that every individual achieves completeness or is complete at all times. It is a technical expression of the earlier sentiment that, for all we know, existence is something toward which all things tend. The process of seeking completeness may be seen as competitive, and the successful determination of some individuals may preclude the successful determination of others.

By introducing the principles so far, I am incrementally building a dipole vocabulary linking the ideas of abstractness, indeterminateness, incomplete natures,

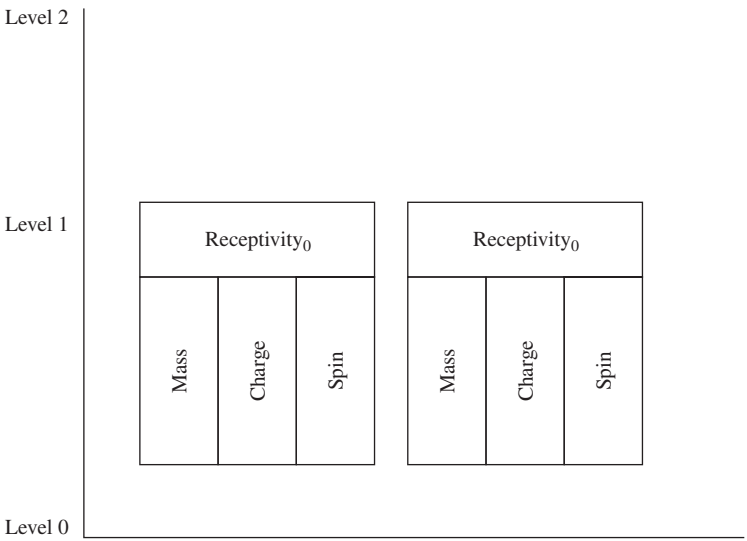


Figure 9.11 Two level-one individuals visualized as elementary particles of some sort. Their states are not represented, but assume that they are indeterminate when considered independently.

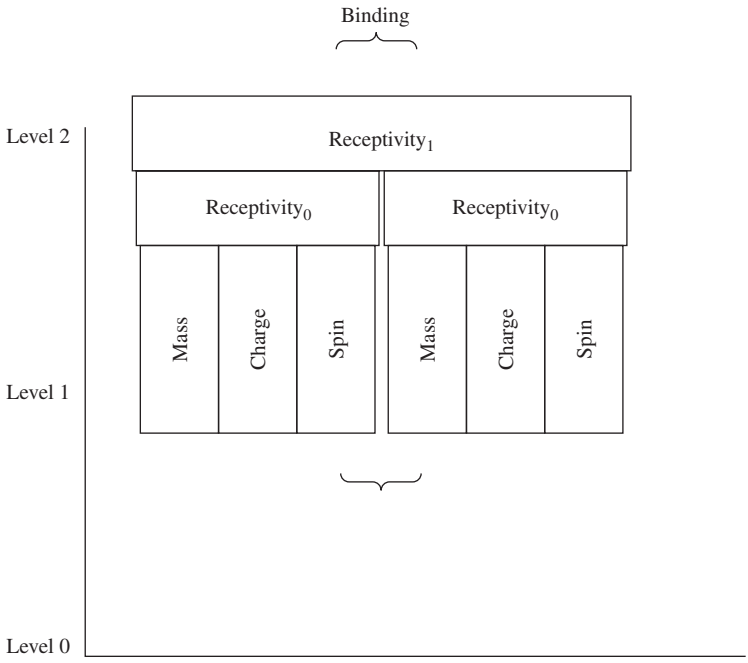


Figure 9.12 The two level-one individuals from figure 9.11 are bound to a common level-one receptivity within which they may become more determinate. This nexus constitutes a level two individual to which the level-one receptivity belongs.

and potentiality on one pole while linking the ideas of concreteness, determinateness, completion, and actualization on the other pole. A concrete event can be seen as the completion of an atemporal process of becoming moving an individual from one pole to the other. Determination is an actualization, a coming into the world, for a bound property complex. Furthermore, in building this vocabulary, I am not only linking the concepts at each pole and contrasting the two poles, but I am also making their application, in principle, a *matter of degree*. In the next chapter I discuss how these things can be a matter of degree when I explore the link between possibility and actuality.

Here is an example of the concepts at work in a Newtonian world. Consider the classical properties of Mass, Charge, and Velocity. The present theory holds that these would be basic effective properties and, therefore, level-zero individuals in the classical world. They would be found as members of effective/receptive complexes, where they share a common receptivity, creating level-one individuals. Let P_1 be such an individual and be represented by the notation [Mass. Charge.Velocity]. P_1 would be a basic particle.

The instance of receptivity binding the instances of Mass, Charge, and Velocity

within P_1 is now complete. Recall transitivity: When an instance of a symmetric receptivity such as P_1*'s* binds two or more effective properties, it becomes part of the completion for all of them, whereas, through the same operation, they become parts of its completion, too. So, for example, in a [Mass.Charge.Velocity] individual such as P_1, Charge is part of the completion of the receptivity, which is itself part of the completion of Mass and Velocity. By transitivity, each of these effective properties likewise becomes part of the completion of the others.

These effective natures, precisely because they are *effective* natures, must share relations of intrinsic compatibility, inclusion, and exclusion with one another. The placement of these restrictive relations has the effect of determining under what conditions an effective property may properly form part of the completion of another effective property, thereby placing restrictions on the copresence of effective properties within a single nexus. On this view, stable particles such as P_1 are those property complexes that contain effective properties with a determinate set of values that are highly compatible or, equivalently, properties with a value set where the values minimally constrain one another, implying that Mass, Charge, and Velocity are in some sense highly compatible properties that form a stable nexus.[10]

Using these ideas, one can give a metaphysical account of what an *immediate causal interaction* is by viewing it as the creation of a level-two individual from one or more incomplete level-one individuals. Let us say that the particle P_1 has a second order property of *acceleration* that is not made determinate by conditions internal to the P_1 nexus. This implies that P_1 is not complete. Let us also say that its acceleration *is* made determinate by these conditions *plus* the magnitude of a certain force F at the region of space occupied by P_1. Resolving the determination problem requires P_1 to receive the constraint associated with this force.

P_1 has a receptivity belonging to it and further constraint may come to P_1 through the receptivity belonging to it. However its receptiveness is only potential until P_1 itself enters into a causal nexus defining its receptive field, i.e., providing the context in which further constraint may be received. There is no problem here. Although P_1 itself is a causal nexus of individuals at one level, that does not preclude it from becoming part of a causal nexus at another level.

P_1's receptive field will consist of other individuals from whom it receives constraint through a shared receptivity, so to realize its potential for having a receptive field, there must be this distinct receptivity binding P_1 into a higher level causal nexus. We use P_1(_) to represent an instance of receptivity bound to P_1's nature. This receptivity is a level-one receptivity *binding* to P_1, and it is distinct from the level-zero receptivity *belonging* to P_1. This irreducible higher level receptivity establishes P_1*'s* receptive context and thereby allows nature to redress incompletenesses in its nature. The other members of this new nexus will constitute what other individuals, if any, are in P_1*'s* receptive field.

To deliver constraint to P_1, the force F must saturate the open slot in P_1(_). If we presumed that the force F's magnitude is not affected by P_1, we would represent asymmetric constraint with the formula $F \Rightarrow P_1$. The nexus $F \Rightarrow P_1$ is a

level-two individual representing the action of the force on that particle at that region of space. P_1 is a level-one individual containing the property of *velocity,* and the new individual makes the second-order property of *acceleration* determinate for P_1. In other words: P_1 is receptive to *F; F* is in the receptive field of P_1. If we presumed that the magnitude of the force also depends to some degree on P_1, we would model the symmetrically connected level-two individual as [P_1. *F*]. This model of direct interaction as the creation of a new level of individual in nature is illustrated in figure 9.13.

We can understand a simplified model of billiard ball causation in a similar way. Imagine that billiard balls are continuously dense spheres with four properties: Mass, Velocity, Shape, and Direction (i.e., each ball is an individual of the form [mass.velocity.shape.direction]). If there are two billiard balls, *B1* and *B2,* with *B1* traveling toward *B2,* one way to understand the causal situation is depicted in figure 9.14.

B1 and *B2* are causal processes, meaning that each temporal stage of the billiard ball shares an asymmetric receptive connection to the previous stage. The single-headed arrows connecting the different temporal stages of the billiard balls represent these asymmetric receptive connections in the figure. Through these asymmetric connections, the immediately earlier stage of a billiard ball constrains the state at the later stage without being constrained in turn (the earlier stage can be seen as in the receptive field of the later stage, but not vice versa). In the first time slice of the figure, *B1* has a certain velocity and direction that are taking it toward *B2*. The collision between *B1* and *B2* creates a natural individual of which they are members and that exists only in time 2.

This new natural individual, the collision, represents a symmetric interaction between *B1* and *B2,* depicted by the box around the billiard balls and the two-

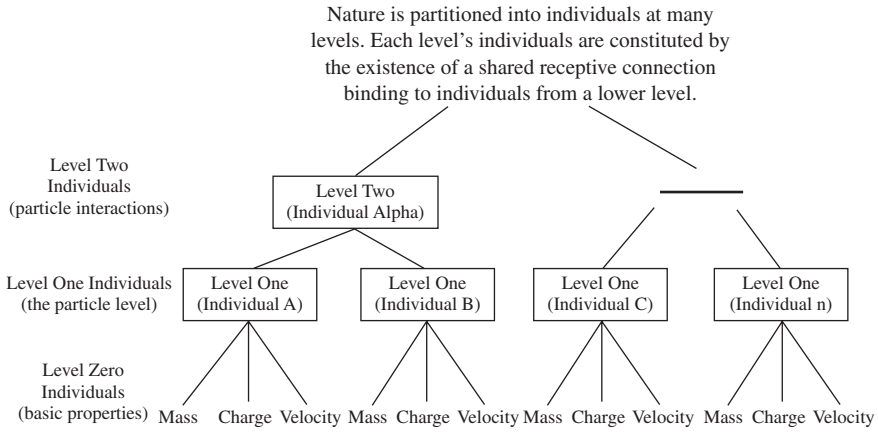


Figure 9.13 Levels of fundamental causal connection in nature may ascend as high as necessary to ensure determinateness.

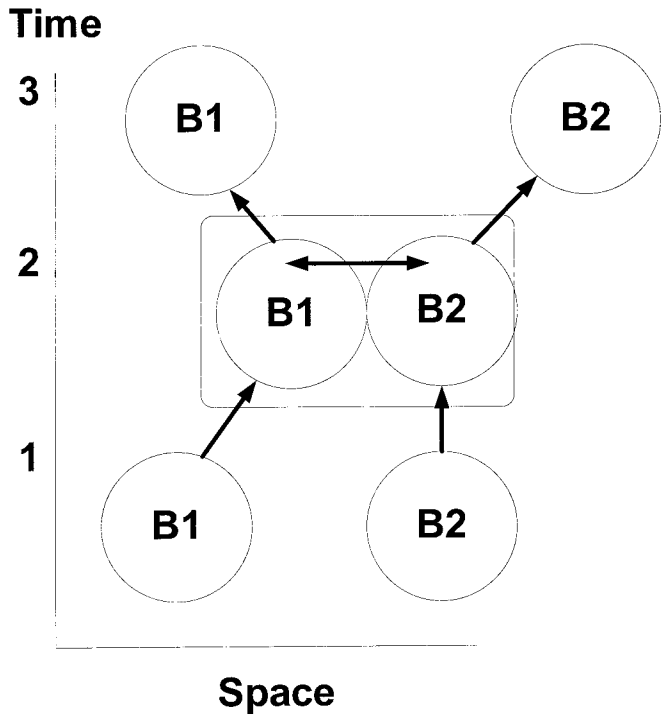


Figure 9.14 Two billiard balls colliding with one another.

headed arrow between them. This adds a new constraint in addition to the asymmetric constraint each ball has to its previous state (*B2,* as well as the earlier stage of *B1,* is in the receptive field of *B1* at time 2). The total situation forces the state of each ball to be compatible both with its own previous state and also with whatever the current state of the other ball becomes. The constraint structure results in a new velocity and direction for each ball, with the consequences of these changes seen in time 3, where the balls are separated but still must take on states compatible with their own previously established states.

This treatment of the relation between the effective properties, the receptive properties, and natural individuation makes sense of some of the traditional views about receptivity. For instance, the medieval conception of God, used by Locke, as an entity that is *above all passive power* is inherited from the theological intuition of God as being intrinsically complete, whereas the created world is somehow inferior to and dependent on God's nature. This situation gets represented straightforwardly by introducing *God* as a complete nature and the world $W(\_)$ as an incomplete nature and postulating an asymmetric binding $God \Rightarrow W$ that represents the asymmetric flow of effective constraint from God's nature to

the world. Furthermore, the intuitive oddness of thinking of receptivity as a kind of *passive power* is removed, as it is more natural to think of it as a kind of *openness* bound to the nature of effective determinables than as a kind of *power*.

In the Newtonian picture of the world, interactions are modeled as level-two individuals consisting of the binding between a particle and a force. This is all the stratification we would ever need in a classical world, but there clearly is no *metaphysical* reason that worlds should be so shallow. So far, there are level-zero individuals. These are the fundamental physical quantities and the fundamental receptivities. There are also level-one individuals. These are the bindings of these physical quantities with level-zero receptive connections to form particles and fields of force. Finally, there are the level-two individuals, and these are the bindings of particles with fundamental forces. This is a clean and simple picture that solves the determination problem quickly and intuitively; however, this classical view of the world is not a correct view of the world. Certain features of quantum mechanics (such as quantum entanglement) at least suggest that the actual world really is more richly structured than this Newtonian picture suggests. This opens up the possibility of a nice inductive definition of natural individual:

> *Natural individual, base case:* Any primitive effective or receptive property is a level-zero natural individual.
> *Natural individual, inductive case:* Any set of natural individuals of level *N* bound into a completed receptive connection constitutes a natural individual of level *N+1*.

This inductive definition allows the world potentially to be a place with a great depth of individuals corresponding to many layers of binding and completion before full determinateness is achieved. Each individual would have an irreducible component, its receptivity, and a set of reducible components, the lower level individuals that are bound by its receptivity. Imagine that there were plastic binders that could turn six six-packs of coke into a thirty-six-pack and other binders that could create two-hundred-and-sixteen-packs from six thirty-six-packs, and so on. Figure 9.13 and figure 9.15 each provide a way to picture such a world, with figure 9.15 emphasizing the irreducible nature of each receptive connection.

Finally, I emphasize that I am introducing the term *natural individuals* as a technical term and that they do not correspond in any direct way to the perceptual and conceptual individuals we speak of in daily life. I even take it to be a substantial empirical question as to whether the individuals within a successful scientific theory are natural individuals in the sense that I have proposed. For example, societies may appear as individuals within sociology, and galaxies may appear as individuals within astronomy, but it does not follow that they are *natural individuals*. The *natural individuals* above level-zero are individuals in virtue of the fact that they have a special, unitary causal nature. They each consist of an irreducible receptive connection through which their components contribute to a set of global constraints on their joint state, and they are capable of having receptive fields of their own. They are "natural" individuals because they have a

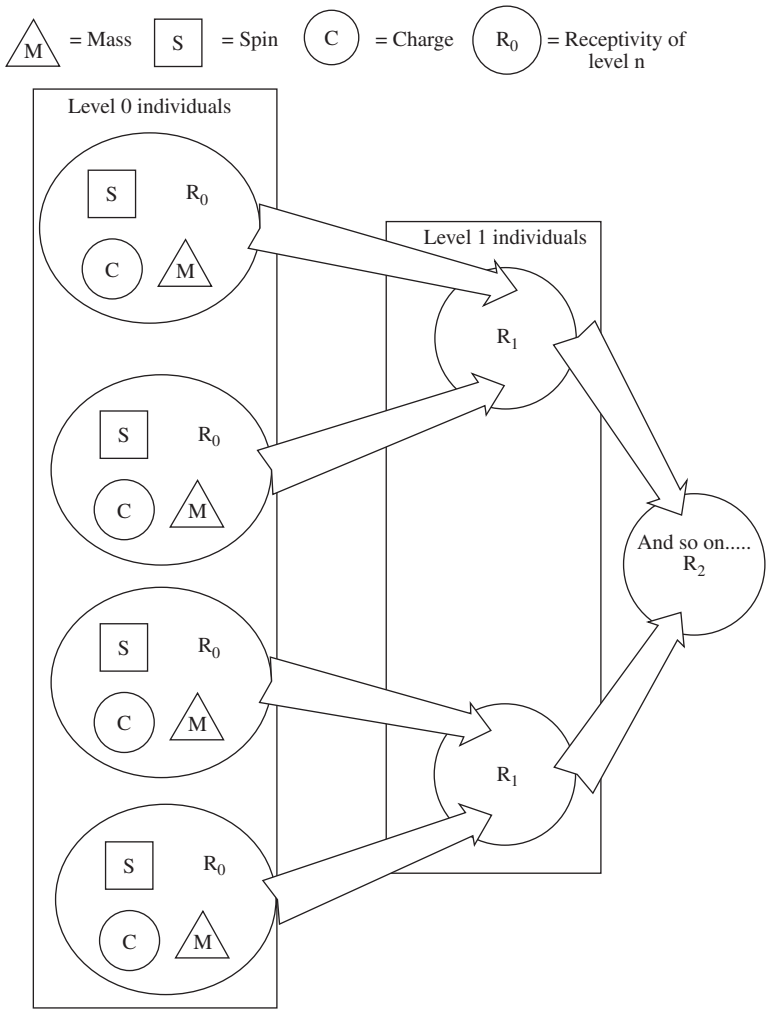


Figure 9.15 Each natural individual at each level has its own unique and fundamental instance of receptivity.

special ontological unity constituted by the merging of their constituents' natures, facilitated by the receptive connections.

9.12 Laws of Emergence for Higher Level Individuals

The model introduced in the preceding section implies that levels of nature are strongly emergent and that each level of nature is a configuration of individuals at the previous level. The configuration consists of a set of irreducible receptive connections, each of which binds a select group of individuals at its own level

into a higher level individual. We can describe the components of the natural order as follows:

1. There are natural individuals. Above level-zero, each natural individual of level n is analyzed into
   - A set S of natural individuals from level $n - 1$. S contains exactly one irreducible receptive connection R of level $n - 1$ and arity k. R is appropriate for binding with the natural individuals at its same level. S also contains a group of size k of other individuals from level $n - 1$.
   - An exhaustive assignment of natural individuals from S to slots of R under a primitive *binding* relationship.
   - A set of independently possible states for the natural individual at level n that results from the binding of R with the other members of S. If this is a singleton set, the natural individual is *determinate*. Otherwise, it is *indeterminate*.
2. There are levels of nature. Each level of nature n is a configuration of natural individuals of level n. Configurations of natural individuals are distinguished by
   - The number and kind of irreducible receptive connections of level $n - 1$ that have emerged and which belong to the individuals of level n. Recall that these receptive connections *bind* individuals at their own level but *belong* to the individual of the higher level that emerges because of the binding.
   - The bindings through which individuals of level $n - 1$ are selected and assigned to the receptive connections of level $n - 1$.
   - The possible states for the world given the emergent constraints associated with the configuration at level n.

The inputs into the configuration of individuals at level n are the natural individuals of level $n - 1$ and their possible states. The possibility of a new level comes from the emergence of receptive connections also of level $n - 1$ able to instantiate new constraints by binding the individuals at level $n - 1$ into level n individuals. The result of the configuration is a new set of individuals, each with their own possible states. The possible states of these new individuals are a selection from the joint independently possible states of their members according to a set of constraints corresponding to their internal relations of inclusion, exclusion, and compatibility, and therefore present a new (smaller) set of possible states for the world. Figure 9.16 illustrates levels of nature filtering possible states of the world.

The key question is: By what rules are the configurations of each level chosen? From a purely combinatorial point of view, for any given level of nature one could construct an enormous number of possible configurations for the next level. If the causal significance view of causation is correct, there must be some way nature chooses one configuration over another. These are laws of emergence for higher level individuals.

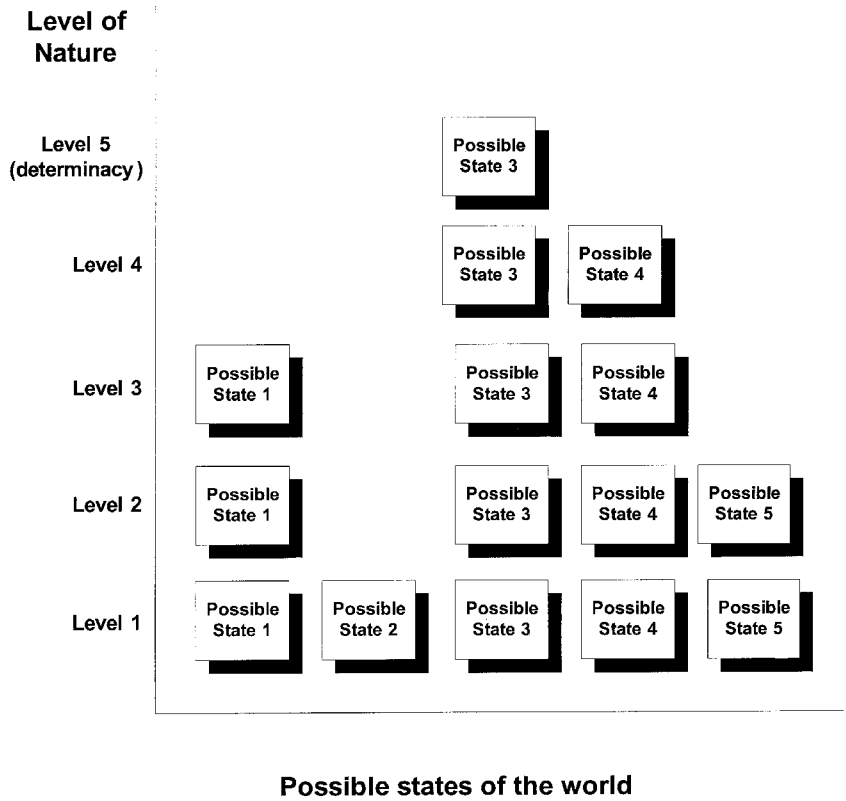


Figure 9.16 At each level of nature the emergent causal constraints filter out some possible states of the world. New levels cease to emerge after the level where the world is in a determinate state, here depicted as level five.

One rule is obvious: There must be indeterminacy in the lower level for a higher level to emerge at all. Without indeterminacy in the lower level, the determination problem is solved, and there is no need for further causation.

If there is indeterminacy to resolve at the lower level, then there is need for further causation. The alternative configurations for the emergent level will have properties of their own, determined by the states of the individuals that constitute them, and it is natural to suppose that the choice among configurations would be a function of their properties. Recall that each constituting individual has a number of independently possible states. In considering how or why some configurations might be preferred over others, reflection suggests two principles of interest to nature that could be relevant:

1. The principle of maximal completeness.
2. The principles of thermodynamics.[11]

We can evaluate the set of independently possible states for a given individual for both its degree of completeness and the level of entropy within it. With regard to entropy, each independently possible state of each individual will have a degree of entropy, and we could measure the entropy for the individual by taking an average of the entropy of its independently possible states. The entropy of the configuration would be a function of the entropy of the individuals within it.

With regards to completeness, the principle of maximal completion says that determinateness indicates completeness. The determinateness of the configuration is a function of the possible joint states of the individuals within it. The fewer possible joint states a configuration allows, the more determinate it is.

Having said all this, I cannot propose a concrete law for the emergence of configurations at higher levels. Yet it seems natural to suppose that the right law might be a function involving nature's dual concerns for maximizing entropy and completeness. That is, given a configuration of individuals at one level, a configuration of individuals at the next level might emerge according to some function of its entropy (as measured by thermodynamics) and its completeness (as measured by determinateness). The precise form of the law could be deterministic (choosing the "best value" along the dimensions) or probabilistic (weighting a probability density function using a measure on the dimensions) and may use both factors or choose one as trumping the other. In any given world, it would be an open question what the precise form of the emergence law(s) would be. This leaves open six possible classes for the laws governing the emergence of higher level individuals:

	Deterministic	*Probabilistic*
Use untropy and completeness together	The configuration emerging at the higher level is governed by a deterministic function attempting to maximize both the entropy and completeness of the chosen configuration, according to some weighted measure.	The configuration emerging at the higher level is governed by a probabilistic function attempting to maximize both the entropy and completeness of the chosen configuration, assigning probabilities according to some weighted measure.
Use completeness alone	The configuration emerging at the higher level is governed by a deterministic function attempting to maximize just the completeness of the chosen configuration.	The configuration emerging at the higher level is governed by a probabilistic function attempting to maximize just the expected completeness of the chosen configuration.
Use entropy alone	The configuration emerging at the higher level is governed by a deterministic function attempting to maximize just the entropy of the chosen configuration.	The configuration emerging at the higher level is governed by a probabilistic function attempting to maximize just the expected entropy of the chosen configuration.

9.13 Summary

I began by arguing that our ordinary notion of causal responsibility is not a purely objective notion. I argued that it rested on an objective core concept involving connections of real constraint between distinct entities, made specific by giving values to a variety of general parameters, and extended by intentional features such as the drawing of figure/ground relations. I called this core notion *causal significance* and presented a theory of it by describing the natures of the different types of causally relevant properties. I called this set of properties the *nomic content* of individuals, arguing that nomic content divided into effective and receptive properties, and I gave a theory of the relations between them. The metaphysical system elaborated in this chapter is a specific articulation of four reasonably intuitive ideas:

1. The world contains effective properties.
2. The world contains receptive properties.
3. Effective and receptive dispositions are categorically linked.
4. A causal nexus is an individual with at least two effective individuals and exactly one receptive connection.

In elaborating these ideas, I developed a view of individuals as pure property complexes by using receptivity as the causal connection and proposing that internal relations between incomplete natures would allow them to mutually complete one another. Effective properties were modeled as determinables that become determinate by conditioning one another. Conditioning is a state in which one effective individual may reduce the potentials of one or more others it is bound to by contributing to the constraints on the nexus of which they are part. Constraints come from intrinsic relations of compatibility, inclusion, and exclusion possessed by effective properties. Within the nexus, each effective property becomes part of the nature of other effective properties through their common binding to an instance of receptivity. It is by becoming part of another property's natural state through a shared receptivity that an effective property may place its constraint on other effective properties. This is one way to elaborate the intuitive notions, and I believe it is reasonable, given the determination problem and our current scientific knowledge. Yet reasonableness is one thing and fruitfulness is another. How far can this elaboration take us in understanding causation and the deep structure of the natural world?

10

A Tutorial on Causal Significance

10.1 Overview of the Tutorial

In chapter 9 I argued that our ordinary notion of causal responsibility was an intentional notion built on an objective core. I called this objective core *causal significance* and identified it with an operation on a space of possibility. The causal significance of a thing is the constraint its presence adds to the space of possible ways the world could be. This chapter explains some key features of causal significance by introducing some working illustrations of its principles. The illustrations take the form of diagrammed situations from a simple physics, with accompanying discussions about consequences of the depicted situations. They divide into two suites. The diagrams in suite 1 illustrate the following key features from a world with only one layer of individuals:

- The character of causal processes.
- A distinction between strong and weak determinism.
- A definition of indeterminism.
- An account of causal counterfactuals.
- Examples of both mediate and immediate interaction.

The diagrams in suite 2 apply the model of causal significance to situations with multiple layers of individuals. They show how to understand:

- The possibility of strongly emergent laws.
- The constitution of higher level processes and individuals.
- An account of epiphenomenal individuals.
- A definition of emergent effective properties.

Finally, the latter part of the chapter introduces some further metaphysical issues surrounding the theory of causal significance. These are issues involving the

nature of possibility, space, time, and the unity of the world. The discussion in the latter part of the chapter points out interesting potential impacts on important areas of metaphysics, and it is mainly a call for more exploration in future work.

10.2 How to Understand the Diagrams

In this chapter I use diagrams of an imaginary physics to explore some models of the causal nexus. The symbols in these diagrams are explained in the following list. I examine a conception of the causal nexus as a collection of natural individuals sharing a common receptivity. All the situations depicted in suites 1 and 2 contain a determinable effective property, which I call *charge,* that may take on one of two values: + (positive) or – (negative).

Guide to the Diagrammatic Language

1. A + or a – represents the instantiation of a value of charge.
2. Every box demarcates a natural individual.
3. A box immediately surrounding a group of +'s or –'s represents a level-one individual and signifies that they are bound together by a receptive connection.
4. A box immediately surrounding a group of level-one individuals represents a level-two individual and signifies that they are bound together by a receptive connection.
5. The ? character in a box represents no commitment to a particular value of charge.
6. Lines with attached beads represent the receptive connections between members within the nexus. These lines may have an arrowhead on one end in place of a bead.
7. Each bead or arrowhead on a line represents one individual bound to the receptive connection.
8. Beaded lines with arrowheads on one end represent asymmetric connections, with the individual pointed to by the arrow unilaterally receiving the constraints placed by the other individual.
9. Beaded lines without arrows represent symmetric connections in which all bound individuals receive constraints placed by all other bound individuals within the connection.

The diagrammatic elements should be read as follows. The two axes represent space and time. The positive and negative charges are represented by + and – symbols placed in relation to space and time. For example, in the first sample diagram (figure 10.1), there are three instances of charge, each at the same placement in space and subsequent to one another in time.

These property instantiations are *level-zero* individuals. I represent the copresence of these individuals within *level-one* individuals by placing a box around the level-zero members of the level- one individual. In figure 10.2, the box sur-

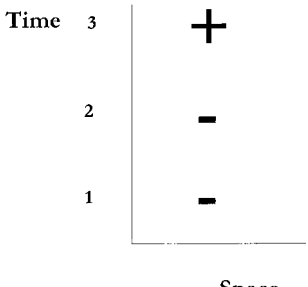


Figure 10.1 An example diagram showing three basic instances of charge instantiated through time.

rounding the negative charges at times 1 and 2 shows both to be within a level-one individual spanning the two moments in time.

Different level-one individuals can share members. In figure 10.3, the negative charge at time 2 is a member of two different level-one individuals. It is bound within one individual spanning times 1 and 2 and also within a second level-one individual spanning times 2 and 3. The diagram shows the overlap between the two level-one individuals as an overlap in the boxes surrounding their members.

Higher level individuals come into being when their members are all bound by a common instance of receptivity. Some instances of receptivity are asymmetric. Lines with arrowheads pointing to the receptive member of an asymmetric connection represent those kinds of receptive connections. The arrow in figure 10.4 represents an asymmetric receptive connection binding the components of the level-one individual spanning times 1 and 2. The individual the arrow points to is the individual that is receiving constraint in the connection. The individual the bead is next to is the individual that is placing the constraint.

An individual *A* is *asymmetrically connected* to *B* just in case *A* is constrained by the effective state of *B*, but not vice versa. In such a connection, *B* has causal significance for *A*. In text, the notation $[I_k \Rightarrow I_j]_{Im}$ represents an asymmetric connection where individual I_k is constraining individual I_j. I_m (sometimes suppressed) names the higher level individual created by their binding. The sub-

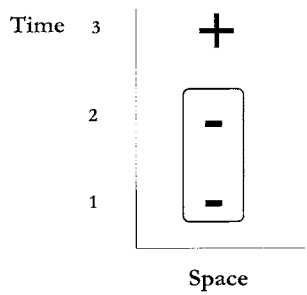


Figure 10.2 An example diagram showing two instances of negative charge bound together inside a level-one individual.

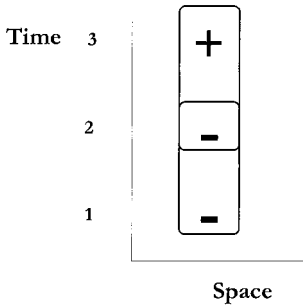


Figure 10.3 An example diagram showing a second level-one individual that shares one of its members with the level-one individual in figure 10.2.

scripts represented by *j, k,* and *i* are numbers such as 1.1, 1.2, 2.1, 2.2, and so forth, where the first number represents the level of nature at which the individual exists and the second number distinguishes the individuals at that level.

Receptive connections can also be symmetric. When a symmetric connection binds individuals, there is no arrowhead. Instead, a beaded line, with a bead next to each member of the connection, is used to represent symmetric connections. Figure 10.5 uses a two-bead line to represent a two-place symmetric receptive connection spanning times 2 and 3.

The diagrams represent symmetric connections using lines with beads on both ends and perhaps in the middle, depending on the number of bound individuals. Within a diagram, the beads are placed in such a way that indicates which individuals the connection binds, and the number of beads represents the carrying capacity of the connection (i.e., the number of "slots"). Figure 10.5 represents the simplest case of a symmetric connection, which is a bidirectional connection. Two individuals, *A* and *B,* are bidirectionally connected just in case a two-place receptive connection exists through which *A* is constrained by the effective state of *B* and *B* is also constrained by the effective state of *A*. I then say that *A* and *B* *share* or are *bound by* a common receptivity.

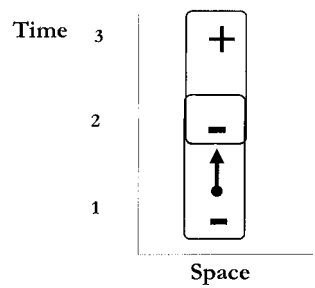


Figure 10.4 An example diagram showing the receptive connection binding the instances of charge within the first of the level-one individuals. The receptive connection is asymmetric. The instance of charge at time slice 2 is depicted as receiving constraint from time slice 1, but not vice versa.

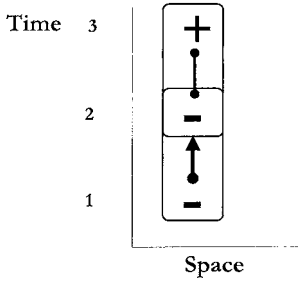


Figure 10.5 An example diagram showing the receptive connection binding the instances of charge within the second level-one individual, as well as the first level-one individual. Unlike the first individual's receptive connection, this individual's receptive connection is depicted as being symmetric, meaning that the instance of charge at time slice 2 constrains the instance at time slice 3 and also receives constraint from it.

In the general case, a collection of *n* individuals *A, B, C* . . . are *symmetrically connected* just in case they share an *n*-place receptive connection through which each individual in the collection can be constrained by the effective state of every other individual in the collection. A completed *n*-place symmetric connection forms an *n*-member nexus of mutually constrained and constraining individuals. These individuals are elements in a simultaneous constraint satisfaction problem presented by the existence of the whole. In text, the notation I use to represent a collection of individuals bound to a common symmetric receptive connection is $[I_1.I_2.I_3...I_j]_{ik}$, where I_1 through I_j name the bound individuals and I_k (sometimes suppressed) names the individual they are bound within.

Receptive connections can stretch across space, as well as time, implying the same for the individuals that they help constitute. Figure 10.6 represents two streams of property instantiations separated in space and an individual at time 3 spanning the spatial distance.

Finally, *charge* in the toy physics will behave differently than it does in our world. More than one instance of charge can occur (i.e., be bound) within an individual, and there is a single causal law restricting its occurrences. The causal law is:

> Each value of charge, + or –, must have an odd number of occurrences in any *natural individual* where that value of charge occurs at all.

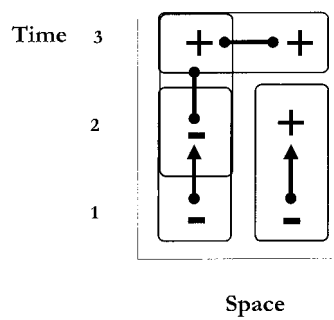


Figure 10.6 An example diagram showing the existence of individuals at other places in space. Notice that there is an individual with a symmetric receptive connection stretching across space in time slice 3. This depicts the fact that receptive connections are not limited to being only temporal connections.

The preceding figures do not conform to the causal law, but the illustrative diagrams that follow do conform.

10.3. Suite 1: Level-One Individuals

Diagram (a). The individuals in diagram (a) (figure 10.7) are causally degenerate individuals, meaning that they do not represent the placement of constraints *between* individuals. Diagram (a) does not properly represent any receptive connections as they have been defined for a world with a legitimate causal character, and I introduce it only because of its purity as an illustration of the idea of constraint and its relation to causal laws. The boxes around the + and – properties indicate that they are each bound within a level-one individual. Each individual's receptivity is a one-place "connection," meaning that it binds just the single instantiation of charge. In text notation, each individual would be represented by a formula such as $[I_j]_{Ik}$, where the name I_j should be replaced by the value of the effective charge that is the bound level-zero individual, and I_k should be replaced by a name for the higher level individual. For example, the two level-one individuals instantiated in time slice 1 should be represented as $[+]_{1.1}$ and $[-]_{1.2}$.

Trivial causal structure. The causal situation in diagram (a) trivially constrains each effective property to be compatible with itself only. The question marks within the boxes at times 2 and 3 represent the fact that those individuals can instantiate either value of charge, + or –, independently of the state of any other individual. This world is operationally equivalent to a Humean world, yet, were it legitimate, it would *not* be a world without real causation: It would have some causal structure, albeit trivial causal structure.

Even though it is counterintuitive to claim that a world like the one depicted in diagram (a) would have causal structure, one can see the putative causal structure in it by trying to imagine a very similar world with a different causal law. Imagine that the law governing the values of charge prohibited *odd* numbers of same

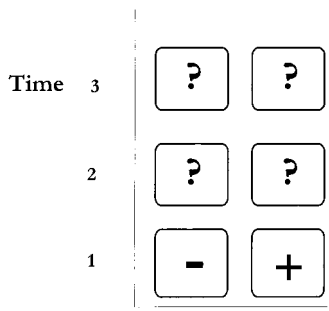


Figure 10.7 An example diagram showing degenerate individuals.

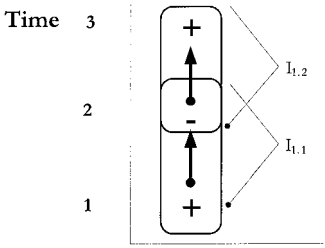
valued instantiations within a nexus instead of prohibiting even-numbered instantiations. Degenerate individuals like those in diagram (a) could not exist because any lone occurrence of a property within a "nexus" would violate the law. Even in an anarchic world like this, it would be a substantive causal matter whether or not there could be trivial individuals of this type.

Diagrams (b) & (c). Diagram (b) (figure 10.8) remedies the degeneracy from diagram (a) by giving the receptive connections temporal depth using asymmetric receptive connections, one binding instances of charge in times 1 and 2 (labeled individual $I_{1.1}$) and one binding instances in times 2 and 3 (labeled individual $I_{1.2}$). Note that the – charge value at time 2 is shared by the two higher level individuals. The result is a *causal process* with two overlapping level-one individuals. These two level-one individuals[1] are represented in text notation as $[+ \Rightarrow -]_{1.1}$ and $[- \Rightarrow +]_{1.2,}$ respectively. In diagram (b) the value of the + effective property at the beginning of the process in time 1 is *fixed* for the rest of the process: It is not receptive to the constraint of any other individual and is thus a given relative to the rest of the process.

Diagram (b) is a good illustration of the three-tiered character of the constraint structure in the causal significance model of causation. On the first tier of constraint, nature must respect the causal laws governing the effective states of individuals when they are bound by a common receptivity. These laws present *universal* constraints applying to every causal nexus, and they state general parameters delimiting the natural possibilities for the effective states of the universe. In the models of this suite, only one law is active, the law prohibiting even numbers of same valued instantiations of charge. In a realistic world, more sophisticated laws would hold and would support more richly structured constraints.

On the second tier of constraint, there is a network of receptive connections. These receptive connections can overlap, binding individuals together in a mesh-like way. The effective property in time 2 of diagram (b), for example, belongs to two individuals simultaneously. That means that its determination is subject to two different constraint structures, one belonging to each individual of which it is a constituent. Therefore, its value must be part of a solution for each set of constraints. This second tier of constraint is very important, as different topologies for the receptive network may exclude different possibilities for the effective states of the world. The world's receptive structure, abstracted from any particular occurrences of effective properties, forms a kind of skeleton that determines a specific set of boundaries within which the causal laws hold.

On the third tier of constraint, any effective determinable whose determinate value is "fixed" relative to the nexii of which it is a member presents a constraint directly to those nexii. The final effective state of an affected nexus must include the fixed value of the relevant effective property at the designated slot. Fixed values will always be on the constraining end of asymmetric connections, and to illustrate the difference between a symmetric and an asymmetric constraint, diagram (c) (figure 10.9) presents another constraint structure, differing



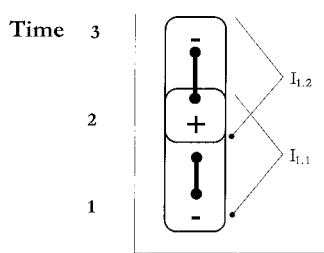
(b) An asymmetrically connected causal process

Figure 10.8 An example diagram showing a causal process. A causal process is a series of individuals containing overlapping receptive connections.

from the one in diagram (b) only by replacing asymmetric connections with symmetric ones. Without the asymmetric connection belonging to individual $I_{1.1}$ from diagram (b), in diagram (c) the first value of charge is no longer fixed, and so the sequence of effective values is free to be different from that in diagram (b).

This three-tiered constraint structure allows for a distinction between (1) an actual state of affairs with all the properties and connections identified and (2) the receptive structure of that state of affairs, that is, only the connections (arrows and beaded line segments) and the lines drawn around individuals but without the particular values of effective properties (+ or –) determined. The receptive skeleton provides the *generic infrastructure* underlying an individual's causal significance within the world. Any particular way of filling in the actual effective properties (e.g., values of charge) creates the *detail* of its causal significance and is legitimate as long as it allows the same skeleton to exist without violating the causal laws.[2]

With these three tiers of constraint laid out, the causal significance view of causation is essentially in place. It represents the objective core of causation in



(c) A symmetrically connected causal process

Figure 10.9 An example diagram showing a causal process where the receptive connections are symmetric. Compare to the causal process in figure 10.8, where the receptive connections were asymmetric.

terms of nature's solving the determination problem by turning it into a *multiple constraint satisfaction problem*. The constraints operate on a space of possibility, each effective property is cast as a variable whose potential values are possible solution values, and the receptive connections themselves emerge as operators on this space of possibilities that creates structure within which these variables can be related to one another.

As examples of the relation between causal significance and constraint satisfaction, the processes in diagrams (b) and (c) illustrate two simple constraint structures. The set of constraints associated with the process in diagram (b) consists of the one general law active in these diagrams and the partial set of causal facts represented by $[+ \Rightarrow I_{0.2}]_{1.1}$ and $[I_{0.2} \Rightarrow I_{0.3}]_{1.2}$. Because the first effective property is fixed at a positive value in time 1, the constraint structure has only one solution, a sequence that oscillates back and forth between + and –, beginning with the fixed + value in time 1.

Because the process in diagram (c) features bidirectional connections, it possesses a different constraint structure than the process in diagram (b) possesses. The symmetric character of the receptive constraints means that no effective values are fixed for this process, so the determinate sequence of –,+, – values shown in the diagram is just one of two possible oscillating sequences satisfying its constraints, the other being +, –, +.

Counterfactuals. The difference of connection type makes a subtle difference to the counterfactuals that hold in the two diagrams. In diagram (b), the causal process is, as a whole, strongly deterministic. This strong determinism shows up in the fact that the constraints of the situation are consistent with only one solution in the space of possibilities. The critical factor is the fixed positive value at time 1, and it is easy to see the value in time 1 as *necessitating* the values at times 2 and 3. Not only is the positive value at time 1 fixed relative to the subsequent causal process, but it is also represented as altogether immune from influence. These facts together raise questions about the proper way to evaluate a counterfactual such as, "If the value at time 3 had been positive, then the value in time 2 would have been negative."

The difficulty is highlighted if we contrast the process in diagram (b) with its near twin in diagram (c). The receptive structure in diagram (c) differs from the one in diagram (b) only by containing symmetric connections that spread constraint bidirectionally across each time slice, with the consequences that (1) the value of no effective property is fixed relative to the whole process and (2) the space of possibilities contains *two* potential solutions to its constraints: $[+.-]_{1.1}$ $[-.+]_{1.2}$ and $[-.+]_{1.1}$ $[+.-]_{1.2}$.

The process in diagram (c) is therefore *indeterministic*, as the causal constraints do not serve to constrain the space of possibilities to a unique sequence. The counterfactual, "If the value at time 3 had been –, then the value at time 2 would have been +," which is analogous to the previous counterfactual, seems to be clearly true. The antecedent corresponds to the first solution in the previous

paragraph, and the consequent is satisfied by that solution. The problem for the first counterfactual, the one about the process in diagram (b), is that there is no analogous truth-maker in the space of possibilities for it, and simply judging it to be true would ignore this seemingly important difference in the causal structure of the situation.

The problem raised by the counterfactual is how to evaluate it for the process in diagram (b) without papering over important facts about the underlying causal structure. We have reasons to resist evaluating it as true, because we would need to let the distribution of receptive constraints shift to find a truth-maker for it, keeping only the causal law constant. Yet only by allowing ourselves to change this deeper second level of causal constraints in diagram (b) can we come to see its counterfactual as true.

Although it is easy to lose sight of the causal connections as part of a constant background, I believe they *are* part of a background that should be held constant when evaluating the counterfactual. The asymmetric process is not the same kind of entity as the symmetric process and should not be shape-shifted into it unless the counterfactual explicitly supposes the shift.

When reviewing diagram (b), the single most important feature of the situation is the absence of a direction of influence from time 3 to time 2. With the value of charge at time 1 fixed relative to its value at time 2, the value at time 2 could not be variable as a function of the value at time 3. In experimental terms, it is not a dependent variable, relative to the value at time 3. Because a natural reading of the counterfactual "If the value at time 3 had been positive, then the value in time 2 would have been negative" suggests influence, I believe that the counterfactual is false when evaluated for diagram (b), despite being true when evaluated for diagram (c).

Nevertheless, the counterfactual has a related true version that respects the direction of influence. The related version is, "For the value at time 3 to have been positive, the value at time 2 would have had to have been negative." I emphasize the difference between these two versions to highlight the directed graph character of the causal model and the realist assumptions underlying it, and so the importance of respecting the direction of the graph in reasoning about it.

Two kinds of determinism. The process in diagram (c) is metaphysically indeterministic, but there is a weaker sense in which one might say it is deterministic. All the forward-looking hypotheticals, such as "If the value in time 1 is +, then the value in time 2 will be –," are true. Additionally, all backward-looking counterfactuals, such as "If the value in time 2 had been –, then the value in time 1 would have been +." are also true. These show an *informational* or *epistemic determinism* shared by both processes in the diagrams. Given background knowledge of the other causal constraints, the information about the effective state of a process at any time slice yields the information about the effective states at all other time slices in the process. In fact, even without the background knowledge of causal constraints, the information about the value along with a natural

law describing the oscillation would make the information about other values available.[3]

The two processes in diagrams (b) and (c), therefore, highlight two different varieties of determinism, a *metaphysical* determinism and an *epistemic* one. The causal constraints in diagram (b) metaphysically determine the process it depicts in the sense that they necessitate its effective character. That process is also informationally deterministic in the sense that knowledge of any one effective state is enough to derive knowledge of the others.

In contrast, the process in diagram (c) is metaphysically *indeterministic* because the three tiers of constraints underdetermine the effective character of that process. Nevertheless, it is still epistemically determined, because the effective character of the process can be determined entirely from an examination of the effective state of any part of it, given knowledge of the causal constraints or the regularities they produce. Metaphysical determinism implies epistemic determinism, but not vice versa. Thus, they seem to be a weak and strong variety of determinism. An important result is this: epistemic determinism does not imply metaphysical determinism.

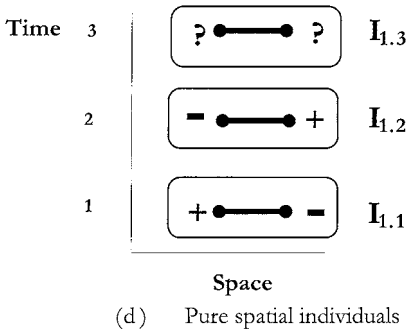
Diagram (d). The receptive connections in diagram (d) are temporally shallow but have spatial breadth.

The purpose of diagram (d) (figure 10.10) is to illustrate that causal significance can be as much of a spatial relation as a temporal one, which makes it significantly different from traditional concepts of causal responsibility. Although each time slice in the diagram is independent of the others, the world is neither Humean nor causally trivial. The receptive constraints in the diagram cut the space of possibilities in half by excluding the instantiation of either two + values of the charge property or two – properties in a time slice. Individuals $I_{1.1}$ and $I_{1.2}$ possess instantiated values to represent the two possibilities that remain, while the question marks in the receptive slots of $I_{1.3}$ emphasize that both possibilities are viable for it, showing its independence from the individuals at the previous times.

Like the temporal processes in diagram (c), the states of the level-one individuals in diagram (d) are metaphysically indeterministic but epistemically deterministic. Each receptive field represents a constraint of the form $[I_j.I_k]$, which has two solutions in the space of possibilities: [+.–] and [–.+]. Thus each field represents an individual whose total state is indeterministically constrained. Nevertheless, if one were to make a measurement of either the right or the left side of the receptive field, one could infer the effective state of the individual on the other side.[4] Note that the model does *not* require individuals bound within a common receptive field to be temporal or spatial *neighbors*. Just as receptive connections can be either spatial or temporal, they can be either local or nonlocal.

Diagram (e). Diagram (e) (figure 10.11) has one field with spatial breadth but no temporal depth, one field with only symmetric temporal depth, and one field with only asymmetric temporal depth.

The purpose of diagram (e) is to illustrate the difference between mediate and



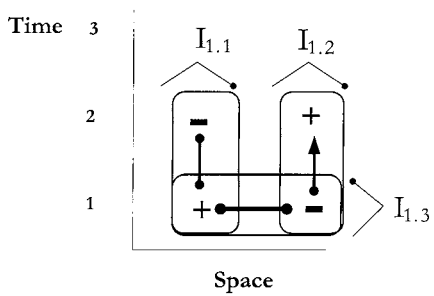
(d) Pure spatial individuals

Figure 10.10 An example diagram showing a series of spatially connected individuals. Notice that there are no connections through time depicted here. As a consequence, this would be a world with causality but no temporal causal relations.

immediate causal connection. Consider the spatial individual $I_{1.3}$ instantiated at time 1. $I_{1.3}$ is just like the spatial individuals from diagram (d). Like them, $I_{1.3}$ must contain one + property and one – property. The temporal process on the left, $I_{1.1}$, is like the symmetric temporal process from diagram (c), and the process on the right, $I_{1.2,}$ is like the asymmetric process from diagram (b).

Notice that $I_{1.3}$ constrains its two constituent values for charge to opposite values in a *direct* sense, meaning that no other receptive connections are involved in mediating it. The only two relevant facts are: (1) they are bound within a nexus to a common receptivity and (2) their effective states must conform to the causal law.

On the other hand, constraints between the spatially separated instances of charge in time 2 are mediated by the overlaps of the level-one individuals on the instances of charge in time 1. In this case, $I_{1.3}$ mediates the relationship between the time 2 constituents of $I_{1.1}$ and $I_{1.2}$. As represented, the time 2 component of $I_{1.2}$ is *responsive* to the corresponding component of $I_{1.1}$ *through* $I_{1.3}$, but they are not bound directly within any higher level individual. The proper phrasing is to say that it is responsive to the value in $I_{1.1}$ but *not* receptive to it. The differ-



(e) A network of spatial and temporal individuals

Figure 10.11 An example diagram showing two temporal individuals connected by a spatial individual containing a member from each of them. Because this spatial individual connects them, their members can interact indirectly, in contrast to the direct interaction that occurs between members within a nexus.

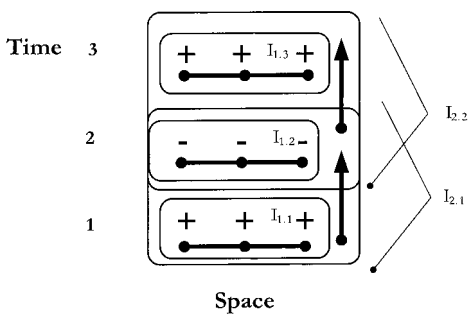
ence between responsive and receptive relations corresponds to the difference between mediate and immediate causal interaction.

10.4 Suite 2: Generalized Higher Level Individuals

The diagrams in this suite illustrate aspects of individuals at level two. Characteristics of individuals at levels higher than level two can be determined from these examples by straightforward extension of the principles illustrated by the level-two individuals. The diagrams in this suite use as the causal law governing level-two individuals the most straightforward extension of the causal law from the previous diagrams. Let $v_+(x)$ and $v_-(x)$ be two functions that accept any individual as an argument and that return the number of + or – properties, respectively, that are instantiated within that individual. The general law for a higher level individual is that each of those functions, when passed an individual as an argument, must yield either zero or an odd number.

Diagram (f). The process in diagram (f) (figure 10.12) shows an asymmetric connection between level-one individuals $I_{1.1}$ and $I_{1.2}$ and another asymmetric connection between $I_{1.2}$ and $I_{1.3}$. Each connection binds two level-one individuals that in turn are binding three instances of charge. The level-two individuals overlap on the level-one individual $I_{1.2}$ in time slice 2. The constraint structure active for these higher level individuals should be prioritized from the bottom up.

- The level-zero individuals are the instances of charge, and they are constrained to have either the value + or the value –.
- The level-one individuals are three-place symmetric individuals constrained by the causal law requiring them to have only odd numbers of + and – values of *charge* as members. Given these constraints, they have only two solutions in the space of possibilities: [–.–.–] and [+.+.+]. These are their independently possible states.



(f) A process with level two individuals

Figure 10.12 An example diagram showing a process consisting of level-two individuals that have level-one individuals as members. The significance of the level-two individuals for the level-one individuals is that they provide a context within which the level-one individuals can constrain and receive constraint from each other as wholes.

- The level-two individuals are two-place individuals with an asymmetric receptivity and are constrained by the causal law for level-two individuals given previously. The effective state of individual $I_{1.1}$ is fixed for level two and is given as $[+.+.+]_{1.1}$.

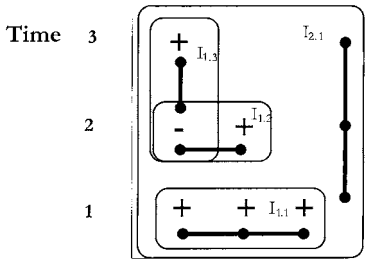
$I_{1.2}$ has its own receptivity belonging to it (the one binding its members) through which it may receive effective constraint from other individuals at its own level. These individuals will constitute its *receptive field*. The concept of a receptive field will turn out to be important when we eventually return to the topic of consciousness, so it is worth understanding it here. In the diagram, $I_{1.2}$ is asymmetrically constrained by $I_{1.1}$ within the higher level individual $I_{2.1}$. $I_{1.1}$ is therefore in the *receptive field* of $I_{1.2}$. In general, the receptive field for an individual consists of those other individuals that may contribute *directly* to constraints on it because they share a receptive connection with it within a higher level individual.

In the preceding, the total set of state constraints for $I_{2.1}$ eliminates [+.+.+] as a realized state for $I_{1.2}$ because [+.+.+] would instantiate an even number of + values within $I_{2.1}$. The removal of [+.+.+] as a possible state for $I_{1.2}$ leaves [–.–.–] as its only remaining possibility. Thus its two independently possible states have been narrowed to one by the presence of the higher level individual, and it has been made *determinate*. This is an example of how a higher level individual can resolve the determination problem by layering constraints.

The analogous pattern of reasoning shows that [+.+.+] is the only viable realized state for $I_{1.3}$. Thus the constraints on the level-two individuals in the diagram force them to be determinate individuals, and their determinateness implies determinate values for individuals at levels one and zero. Notice that there is a kind of teleology here: The higher level individual achieves determinateness for itself, and its achievement requires determinate values for the lower level individuals. There are basically two kinds of determination at work here. There is the bottom-up determination through which the states of the lower-level individuals are the *material causes* of the states of the higher level individual. They constitute its concreteness. There is also a top-down determination through which the states of the higher level individual act as a kind of *final cause* for the lower level individuals in the sense that the achievement of the determinate state of the higher level individual constitutes a reason or purpose for the determination of the lower level individuals.

Diagram (g). The purpose of diagram (g) (figure 10.13) is to show the connection between the constraint character of causal significance and the role of higher level individuals as possibility filters on the lower levels of nature. In diagram (g), the existence of the level-two individual $I_{2.1}$ makes $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$—the level-one individuals bound within it—directly and symmetrically receptive to each other.

Looking at diagram (g), how do we determine the possible states for $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$ in the depicted situation? Obviously, the *internal* constraints on the states



(g) Emergence within level two individuals

Figure 10.13 An example diagram showing a symmetrically connected level-two individual whose members already share some overlapping connections among themselves.

of $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$ are still active. The candidate individual states compatible with their own internal constraints constitute a set of *potentialities* for each of them: These are their independently possible states. The combinations of those solutions constitute possible joint states for the three level-one individuals, and the space of possible joint states is the possibility space that $I_{2.1}$ further constrains. To illustrate by example, the prior possibility space operated on by $I_{2.1}$ is:

	$I_{1.1}$	$I_{1.2}$	$I_{1.3}$
Possibility 1	$[+.+.+]_{I1.1}$	$[-.+]_{I1.2}$	$[-.+]_{I1.3}$
Possibility 2	$[-.-.-]_{I1.1}$	$[+.-]_{I1.2}$	$[+.-]_{I1.3}$
Possibility 3	$[+.+.+]_{I1.1}$	$[+.-]_{I1.2}$	$[+.-]_{I1.3}$
Possibility 4	$[-.-.-]_{I1.1}$	$[-.+]_{I1.2}$	$[-.+]_{I1.3}$

Read each row of the table as constituting an ordered triple, and interpret each triple as a member of the prior possibility space given to $I_{2.1}$. Notice that possibilities 3 and 4 would instantiate an even number of +'s or –'s within $I_{2.1}$, thus violating the causal law. (In checking this, remember that both $I_{1.2}$ and $I_{1.3}$ share their first element, so do not double count its value in the table.) When they become members of $I_{2.1}$, the possibility space for their joint states is thus shrunk in half, because their binding to the common receptivity within $I_{2.1}$ excludes possibilities 3 and 4. The presence of $I_{2.1}$ narrows the possibilities for the joint instantiations of the level-one individuals to:

	$I_{1.1}$	$I_{1.2}$	$I_{1.3}$
Possibility 1	$[+.+.+]_{I1.1}$	$[-.+]_{I1.2}$	$[-.+]_{I1.3}$
Possibility 2	$[-.-.-]_{I1.1}$	$[+.-]_{I1.2}$	$[+.-]_{I1.3}$

Consequently, the existence of the higher level individual has made the world more determinate. By recognizing their autonomous role in resolving the deter-

mination problem, we can see how the existence of layers of higher level individuals opens up a more general way to view what effective properties *are*. We can do this best by contrasting them with effective *states*. An individual's *effective state* is an ordered state, consisting of the occurrence of its constituent's effective properties in particular slots. For instance, imagine a smaller nexus like $I_{1.2}$ that has two possible states, [+.–] and [–.+]. These are the individual's independently possible effective states. It is because ordering matters that [+.–] and [–.+] are different effective states.

An individual's effective state depends on the occurrences of the effective properties of its constituents, but the effective *properties* it has should be defined not by its relation to its constituents but by the way it contributes to constraints on other individuals at its own level. For consistency, we should always identify effective properties with the contributions an individual's effective states make to the constraints placed on the state of a nexus. That is the way we did it at the lower levels, and it is what we should do at the higher levels.

In an individual like $I_{1.2}$, each potential state would add one + and one – to a nexus within which the individual became bound, so within a higher level individual these two states would make the same contribution to the overall constraint structure on the nexus. We can characterize this concept in terms of information by noting that the constituent difference between the two effective states [+.–] and [–.+] is not a difference that makes a difference, because each would represent exactly the same constraint within a higher level individual containing it. Thus, by our definition, although they are different effective *states,* they instantiate the same effective *property* within individuals, the distinction between the two states collapsing as they move through the higher level filter of constraint placement. Effective properties, therefore, come from differences in the effective states of individuals that make a difference to the constraint structure of a nexus.

Strongly emergent laws. Effective properties are thus informational in character, referencing differences they make at their own level of nature, rather than aggregative or constitutional properties referencing the internal structure of the individual to which they belong. Once we begin identifying effective properties with contributions to the constraint structures within possible nexii, the policy I adopted in these examples of establishing higher level laws that are simple continuations of the laws governing lower levels stands out as a convenience. The receptive connection (the nonreductive aspect of each higher level individual) is acting as an operator on a space of possibilities presented by the effective individuals it binds (its reductive components). This circumstance makes it easy to imagine semi-independent laws governing the constraining power of effective states at different levels and mapping them onto unpredictable effective properties.

The emergent laws at each level only have to be *consistent* with the laws governing the effective properties at lower levels and do not have to be *entailed* by them. Such laws would be *strongly emergent*, meaning that they would be funda-

mental laws operating on individuals above the ground level of nature. For example, laws governing the level-two individuals could, consistently with the requirement that only an odd number of any property value can instantiate within level-one individuals, require that only even numbers of any type of property value instantiate within level-two individuals.

Additionally, the forms of these causal laws could quantify over individuated subsets of the effective states of the individuals they govern, and they could describe the effective significance of each quantified subset within a nexus. Causal laws having that form would effectively be providing a coarse-grained individuation of these subsets into new effective properties and a specification of the laws of interaction for these new effective properties. The result of such laws would be *strongly emergent* effective properties.

For example, imagine that the laws governing the level two individuals quantified over property pairs rather than the individual occurrences of + and -. A possible set of laws might dictate matching between pair types within the nexus in the following way,

1. Every occurrence of a […+.-…] sequence within a level two individual must be matched by the occurrence of a second […+.-…] sequence within the individual.
2. The occurrence of any other of the three possible pair sequences […+.+.],[…-.-...], or […-.+...] within a level two individual must be matched by an occurrence of the pair sequence […+.+…] within the individual.
3. No occurrence of + or – may occur within an individual unless it is part of a pair sequence.

Under this regime of causal laws for level two individuals, if two level one individuals of the form [?.?] became bound within an asymmetric level two individual, the total constraint structure would force them each to become determinate with the result, $[[+.-]_{1.1} \rightarrow [+.-]_{1.2}]_{2.1}$. We can understand this result by following the levels of constraint from the bottom up. Looking first at the level one individuals of the form [?.?] that are members of the level two individual, we know that the prior possibility space for them consists of just [-.+] and [+.-], as the causal law mandating odd numbers of property occurrences active for level one individuals precludes states [+.+] and [-.-]. It follows that the prior possibility space for the level two individual consists of the four possible joint states,

1. $[[+.-]_{1.1} \rightarrow [+.-]_{1.2}]_{2.1}$
2. $[[+.-]_{1.1} \rightarrow [-.+]_{1.2}]_{2.1}$
3. $[[-.+]_{1.1} \rightarrow [+.-]_{1.2}]_{2.1}$
4. $[[-.+]_{1.1} \rightarrow [-.+]_{1.2}]_{2.1}$

However, of these four prior possibilities, only possibility one is compatible with the causal laws laid out above. Possibilities 2-4 are all precluded because the pair sequences are not appropriately matched. In these circumstances, the

strongly emergent causal law allows the pair sequences themselves to act as effective properties within the higher level individual, and there are only two such properties because the effective states [+.+], [-.+], and [-.-] form an equivalence class with respect to the constraint they present within nexii. These two effective properties, defined over the constraints presented by pair sequences within individuals effective states, are strongly emergent. As a side note it is worth observing that in this circumstance the presence of the higher level individual forces determinateness on otherwise indeterminate lower level individuals.

This characterization can be extended in a natural way to explain how a higher level individual may have multiple emergent effective properties. To extend the analysis, note that complex individuals may have effective states that are variable along a large number of dimensions, not just two as in the example of [+.–] and [–.+]. Consider an individual whose effective state is given by a very long formula [+.–.–.–.+.–.+……+], which might consist of thousands or millions of constituents. Under the influence of specific kinds of causal laws, an individual could exhibit multiple emergent effective properties if causal laws had the result that the variances within subsets of this sequence have systematic consequences on the constraints it places on other individuals. Different subsets could create distinct equivalence classes corresponding to distinct emergent effective properties, and the number of such subsets would be the dimensionality of the individual's effectiveness. A very complex high level individual, such as perhaps a human brain, could have an effective state consisting of millions or even billions of dimensions and thus could produce a tremendous number of effective properties associated with a single individual.

No paradoxes would result from such strongly emergent laws, nor are there any a priori reasons for believing that they are any more or less unlikely than other kinds of laws. The key feature of this model that allows for such an easy explanation is that a consistency requirement among laws at different levels is easy to enforce: The system builds consistency in from the ground up. The lower level individuals present the domain of prior possibilities to the higher level individuals. Because the lower levels already constrict the domain presented to the higher level individuals according to their own laws, the restrictions belonging to the lower level interactions will always be *already present* at the higher levels. Therefore, the scheme necessarily produces interlevel consistency. Despite making a difference to the behavior of systems, the operation of emergent laws of this sort would not be detectable through a violation of the laws governing the behavior of lower level individuals. They could be present but invisible (or nearly invisible) in their operation: They would show up as mere noise or randomness at the lower levels.

Emergent properties. Whether or not a world possesses emergent laws of the sort described previously, it can possess emergent effective properties. An emergent effective property is a multiply realizable contribution to the constraint structure on possible nexii, as it will now be defined. A *multiply realizable con-*

tribution to the constraint structure on a nexus (1) is a contribution to the constraints on the state of a higher level individual and (2) can potentially be placed by two or more kinds of effective states, or subsets of effective states, possessed by lower level individuals.

Here is a more formal way to think of these emergent effective properties. I say that potential states of an individual I_k possess a *prior* difference to one another if they would be different effective states of I_k (i.e., they would instantiate different constituent structures for I_k). So $[+.-]_{Ik}$ and $[-.+]_{Ik}$, which instantiate different effective states, possess a *prior difference* to one another.

States with prior differences can support the emergence of singular effective properties. Think of every receptive connection *R* as realizing a mathematical function on the vector of individuals it binds. I call the domain of this function the *prior possibility space* presented to *R,* and it is represented by a set of vectors $<I_1, I_2,..., I_j>$, where the I_k's making up the vector stand in for the independently possible effective states of the individuals bound by I. This defines the domain of the function *R*: $<I_1 \times I_2 \times ..., \times I_j>$.

Each vector of the prior possibility space contains a possible *joint state* of the members, considered independently of *R.* That is, each vector represents a possible combination of effective states for the members bound by *R,* as in the example of [+.–] and [–.+] discussed earlier. In the simplest case, the prior possibility space will just consist of the rows in the Cartesian product $I_1 \times I_2 \times ... \times I_j$, where the values of I_1 through I_j are their independently determined prior possibilities.[5]

The range of the function is the *power set* of the prior possibility space that is its domain. The function

$$R: <I_1 \text{ X } I_2 \text{ X...}, \text{ X } I_j> \rightarrow 2^{\{I1 \times I2 \times \ ... \times Ij\}}$$

has the effect of mapping the domain into a subset of itself. It maps allowed vector states onto themselves and disallowed vector states onto the empty set, yielding a set of independently possible effective states for the higher level individual. What is allowed or disallowed depends on the general laws pertinent to nexii of its sort. The allowed states that remain are all the vector states from the original domain that are still possible, given the presence of the higher level individual. Unless the higher level individual is epiphenomenal with respect to its constituent low-level individuals,[6] this remainder will be a *proper* subset of the original domain.

We can use this conception to individuate the possible effective properties that an individual I_k may realize within a higher level individual I_{k+1}. This will capture the *posterior* differences in the effective states of I_{k+1}'s constituents. Let s_1 and s_2 be two effective states for I_k that possess a prior difference, and let tuples such as $<...s_1...>$ and $<...s_2...>$ represent their occurrence(s) in the prior possibility space presented to I_{k+1}. We will say that $<...s_1...>$ and $<...s_2...>$ are *counterparts* of each other, just in case they are exactly the same length, and s_1 and s_2 occur in exactly the same places in their respective tuples. We will say that $<...s_1...>$ and $<...s_2...>$ are *exact counterparts,* just in case they are counterparts,

and they differ in no way except these occurrences of s_1 and s_2. S_1 and s_2 instantiate the *same* posterior effective state within I_{k+1} just in case s_1 and s_2 are interchangeable within I_{k+1}. This means that:

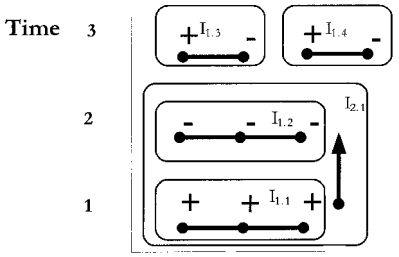
1. Every <...s_1...> has an <...s_2...> counterpart in the prior possibility space, and vice versa.
2. Whenever the function R that represents I_{k+1}'s receptivity maps a tuple <...i_1...> from the prior possibility space onto the empty set, it also maps an <...s_2...> onto the empty set, where that <...s_1...> and <...s_2...> are exact counterparts.
3. Whenever the function that represents I_{k+1}'s receptivity maps a tuple <...s_1...> from the prior possibility space onto <...s_1...>, it also maps an <...s_2...> onto <...s_2...>, where that <...s_1...> and <...s_2...> are exact counterparts.

The preceding definition can be unpacked as follows. S_1 and s_2 are distinct effective states which instantiate the same effective property *within* I_j just in case it is always possible to exchange one state for the other without changing the set of possible states available to any of the other individuals within I_{k+1}. When the possibility of such interchange is the case, the prior differences in I_k's states, s_1 and s_2, are differences that do not make a difference within I_{k+1}. They are informationally irrelevant within I_{k+1}, as they make the same contribution to the constraint structure on the other individuals. Given the concept of an effective property in the causal significance model, only the informationally relevant features of I_k's states present the true effective properties of I_k within I_{k+1}.

States such as s_1 and s_2 are the possible *realization bases* for a higher level effective property that has emerged. One important consequence of this is that the receptive connections of high-level individuals may actually bring irreducible effective properties into existence. Finally, s_1 and s_2 realize the same effective properties *tout court* just in case they place the same effective constraints within all possible I_{k+1}'s. Applied to our example, these definitions have the consequence that state types such as $[+.-]_{ik}$ and $[-.+]_{ik}$, although they share a prior difference, actually realize the same effective property *tout court*.

Diagram (h). Diagram (h) (figure 10.14) illustrates how changes at lower levels may prevent the continued existence of a higher level individual.

In diagram (h), the level-one individuals $I_{1.1}$ and $I_{1.2}$ can support the existence of the level-two individual, $I_{2.1}$, because their prior possibility space contains a solution for the constraint it presents. In the formal sense outlined previously, this means that the function representing the receptivity of $I_{2.1}$ maps the prior possibility space for $i_{1.1}$ and $i_{1.2}$ onto a nonempty set. The relations between lower level individuals may change through time, and it is easy to imagine that subsequent arrangements may not support the kinds of higher level individuals supported by earlier arrangements. These kinds of changes may prevent a higher level process from continuing, and that is what diagram (h) depicts.



(h) A level two individual that cannot maintain itself

Figure 10.14 An example diagram showing how independently developing conditions at a lower level can stop a process from forming or continuing.

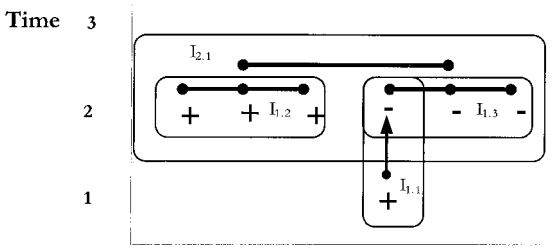
In the diagram, a new level-zero individual appears in time 3, breaking the level-one triple bonding from times 1 and 2 into two double bindings. If an analog to the individual $I_{2.1}$ could be instantiated to bind $I_{1.2}$, $I_{1.3}$, and $I_{1.4}$ (i.e., $[i_{1.2}.i_{1.3}.i_{1.4}]_{I2.2}$), it would be reasonable to see this as a continuation of the high-level process begun as $I_{2.1}$. Unfortunately, no solution exists for $[i_{1.2}.i_{1.3}.i_{1.4}]_{I2.2}$, and so no continuation of $I_{2.1}$ is possible. Formally, the receptive connection of the proposed individual, $I_{2.2}$, would map the proposed prior possibility space onto the empty set. In the sense that processes are de facto individuals, a high level individual ceases to exist.

Diagrams (i) & (j). Diagrams (i) and (j) (figures 10.15 and 10.16) simply represent the two opposite ends of the spectrum of significance for a higher level individual. In diagram (i), the presence the level-two individual $I_{2.1}$ forces a unique solution for the instantiations of the basic effective properties within $I_{1.2}$. Diagram (i) represents a case in which the existence of a higher level individual makes what would otherwise be an indeterministic world into a strongly deterministic one by filtering all but one joint state for its constituents from their prior possibility space. Diagram (j) represents an epiphenomenal higher level individual, $I_{2.1,}$ in that sense that its presence does not constrain the prior possibility space given by its constituents at all.

10.5 Possibility and Actuality

A world in which things have causal significance is a *mesh.* Its deep ontological structure is one of interlocking, overlayed causal nexii, each a natural individual, ordered hierarchically into levels and ordered horizontally across levels, as depicted in figure 10.17.

An individual's nomic content carries its causal significance within the mesh. We can analyze nomic content into two components: the individual's place in a structure of shared receptive connections and the possible effective properties its effective states may realize. Through the actualization of its nomic content, an in-

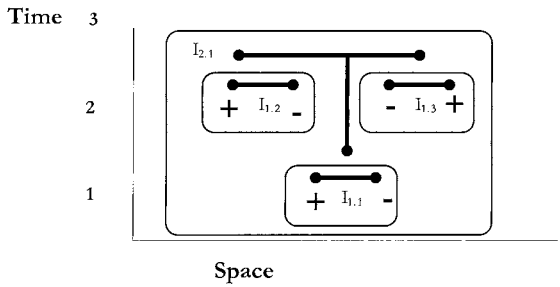


(i) A level two individual that necessitates a unique state for a level one individual

Figure 10.15 An example diagram showing an level two individual creating the conditions for a determinate state in a level one individual.

dividual's presence exhibits direct and indirect constraints on the possible states of other individuals in the mesh. Just as in a mesh, no link can be isolated from neighboring links, with effects of any manipulation possibly propagating in all directions. Causal significance therefore invokes an operation on a space of possibilities.

Conceiving of causation in this manner raises a contentious issue, because there is no explanation here unless *a possibility space objectively exists for causation to constrain.* Some form of robust realism about possibility appears to be an ontological commitment of this view of causation, and, I believe, of any adequate view of causation. The realism needed here is not a Lewisian realism in which all possibilities have the same kind of existence and actuality is only an indexical. Rather, it is a realism in which there are truly different modes of existence, the possible and the actual, and an internal connection and movement of becoming between them. I call views along these lines *abstract modal realism* to



(j) A level two individual that makes no difference to the state of any level one individuals

Figure 10.16 An example diagram showing an epiphenomenal individual.

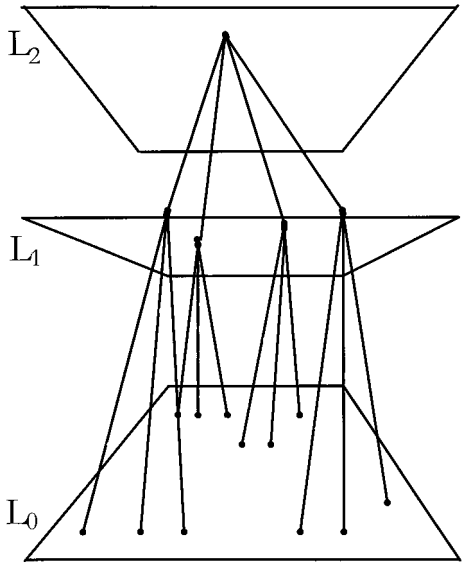


Figure 10.17 This diagram represents the strong emergence of each level from the binding of the individuals at the level below.

contrast it with David Lewis' *concrete modal realism.* Notice that I do not use the word "abstract" to denote unreality or "merely intensional" existence. I discuss abstractness in more detail and produce a model of it later in this chapter.

Could the world be so metaphysically thin that only the actual existed? I argue that *any* adequate view of causation must commit to a version of abstract modal realism because causation has no work to do unless there exist real alternatives to actuality. If there were not any real alternatives to actuality, the existence of causal production and constraint would be illusory. Because the work causal powers perform is essential to them, then, plausibly, they could not really exist within an actualist metaphysics. The formal argument for this conclusion shows that the following two positions:

1. An actualism in which possibilities are fictional constructs
2. The position that some real power of causal production exists

are inconsistent with one another under the usual realist accounts of what it means to be a causal power. Let *C* be any productive cause. To a first approximation, this means that *C* raises the probability that some event *e* will occur relative to some background conditions. In the limit case, *C* by itself is sufficient to make *e* occur, and the background conditions do not matter. The following argument does not apply to the limit case, but I am assuming that a theory of causation covering only the limit case is not adequate.

Assume that positions (1) and (2) are both true. The concept of *C* being a productive cause makes two appeals to possibility. The first appeal is to the probables that do not become actual but might have. The second appeal is in the im-

plication that the probabilities associated with these probables would have been different had *C* not existed. If we are abstract modal realists about these possibilities, we contradict assumption (1). Therefore, the possibilities appealed to must be treated as fictional objects: explanatory fictions, perhaps. Because both the probables and the alternative probability distributions are merely explanatory fictions, the causal power is idle.[7] This contradicts assumption (2). Both cases lead to a contradiction. Therefore, if we agree that an adequate theory of causation must cover the nondeterministic cases, we must conclude either that the power of production is idle, contradicting realism about causation, or that we live in a world with real, unactualized possibilities, contradicting actualism.

The idea of causal constraint, not causal production, is at the core of the theory of causal significance. This difference in emphasis makes mutual causal significance, rather than asymmetric causal responsibility, the fundamental notion. I could easily construct an argument parallel to the preceding that shows that the theory of causal significance is equally committed to abstract modal realism about the space of possibilities. Because any realist theory of causation will build itself on one of these two concepts, it seems that any realist theory of causation will be committed to abstract modal realism about possibility.

By combining the preceding argument that realism about causation implies abstract modal realism about possibility with the arguments for realism about causation given earlier, we arrive at a strong composite argument against actualism. The argument against actualism goes like this: If possibilities were only fictions, then causal realism would be false. If causal realism were false, then, as argued against the Humean, there likely could be no unified world, and we could have no knowledge of the world. Therefore, causal realism is true, possibilities are not mere fictions, and actualism is false. Causation is an operator on a space of real possibilities.

We now have a picture of the *natural* world as a mesh of causal nexii, each nexus carrying causal significance for the rest. The final *metaphysical* picture is one in which the actual world rises in relief against a metaphysical background of possibility: that metaphorical jewel floating in a heaven of transparent possibility. The attempt to more deeply understand nature's connection to its metaphysical background of possibility presents a rich vein of puzzles. In the end, one cannot pretend to understand the natural world fully without recognizing and understanding its relations to its seemingly nonnatural background of possibility.

Reflections. Although these conclusions are strange, they are reassuringly consonant with the character of modern quantum mechanics, and I think this is an interesting picture worth investigating further. I investigate it subsequently, engaging in some very speculative metaphysics. I do it in the spirit of telling a story, telling a tale that I hope may inspire others to do better than I can. My main intents are to provoke thinking and speculation about these very deep matters and to do so in a way that is ontologically bolder, and I think more substantial, than the deflationary approaches that dominated philosophy in the twentieth century.

The speculation I engage in, although provocative or inspiring for some, might be too "metaphysical" for some others. It is likely to leave everyone feeling a little dizzy, as if we left the ground too long ago and too far behind. I believe that, while reading what follows, one may ameliorate some of the pie-in-the-sky feeling it may produce by remembering that the story is mostly systematizing a variety of elements that we already have independent, rational grounds for introducing.

In particular, we already have rational grounds for being realists about causation; for being abstract modal realists about possibility as a consequence of our realism about causation; for believing that facts about causation might allow us to explain the direction of time; for thinking that space and time are inseparable and relative; for introducing a story about the receptive and effective aspects of the causal nexus; and for wanting a story that presents a unified treatment of the natural world and its nonnatural background of possibility. All these considerations are rationally motivated.

Once rational inquiry has taken us this far, we cannot balk at speculations on the grounds that they are too metaphysical. *Some* story has to be told about such things as the direction of time, the unity of the world, and how the natural world is connected to its nonnatural background. It would be strange if such a story did *not* seem a bit strange when viewed from the ordinary, commonsense perspective from which we started our inquiry.

I believe things are always this way in science and philosophy: Common experience presents various phenomena and a need to explain them in an integrated way, and the end of that inquiry leads us to a terrain that looks startlingly different from the one in which we started. Although every new terrain should be critically examined, criticizing any of them simply for being disorienting is unwarranted.

Remember, also, that the following story is tentative and illustrative, intended merely to point in a direction of possible inquiry. Whether this story turns out to be the best one to adopt matters much less than the idea that there is a worthwhile story space for us to explore and the example it presents of one story in the space of possible stories.

Further, some surprising, but potentially fruitful, empirical consequences of the metaphysical speculation may be noted. One very interesting consequence that emerges from the discussion so far is that nothing in the theory of causal significance requires every individual at every level to end up with a determinate effective state. Although we should assume that individuals tend toward determinateness (via the principle of maximal completeness), if the mesh indeterministically constrains an individual, then that individual, consistently with everything that we have said, could hold onto some or all of its initial potentialities. This conclusion has a surprising implication: Worlds could exist in which higher level individuals achieved a level of determinateness greater than that achieved at lower levels. It follows that a person, for example, could be in a perfectly determinate physical state even if some or all of his or her atoms were themselves in indeterminate states.

How could that happen? Recall that the effective properties of an individual are identical with the contributions to constraints placed by its effective states within higher level individuals. The higher level individual's receptive connection may instantiate a many-to-one realization function mapping many distinct lower level effective states onto a single higher level effective property (i.e., constraint). It follows that an individual may have a determinate effective property realized by an indeterminate effective state.

Furthermore, the effective state of the higher level individual is determinate just in case the values of the effective properties of its members are determinate. This concept of the relations between effective properties and effective states at different levels allows for the logical possibility that the effective state of a high-level individual may become fixed and definite *even though the effective states of its constituents remain indefinite.*

This could happen as follows. Imagine that an effective state *S* of a high-level individual is determinate and that this effective state is realizable by either of two states, *s*1 or *s*2, of its constituents. The actuality, the *determinateness*, of the higher level state *S* is compatible with its constituents remaining in the indefinite state: *s1 or s2*. This kind of indefinite disjunctive state becomes easier to grasp once one realizes that a disjunctive state such as *s1 or s2* is logically equivalent to the *conjunctive* state: *potentially s*1 *and potentially s*2 *and not potentially any other state.* Because we are forced into being abstract modal realists about possibility anyway, nothing is inherently objectionable about a conjunction of potentials.

The moral here is deep and deserves more reflection. The indeterminate state of an individual is equivalent to a definite state of that individual understood as a pluralistic selection from its space of potentialities. The example teaches us to refrain from treating individuals as just actualities and to pay attention to their deep roots in the metaphysical background. Everything in nature is thoroughly modal, and the complete essence of an individual stretches along a path from possibility to actuality. Although actuality and possibility are distinct, they are not separable. In the extreme case, it is conceivable that the entire high-level history of a world could enter a perfectly definite series of states even though no lower level individual *ever* possessed a definite effective state.

Even in the face of this possible indeterminacy, the states of higher level individuals continue to logically supervene on the states of lower level individuals. The facts that the lower level individuals are in the indeterminate state *s1 or s2* and that *s*1 and *s*2 realize the same high-level effective state *S* logically determines that the high-level individual is in state *S*. What other state could the higher level individual be in?

The phenomenon does not undermine the ability of low-level states to be the material cause of high-level states but the necessity that they have to be determinate in order to do so. This result is extraordinarily counterintuitive, but, despite this usurpation of common sense, it is a straightforward logical consequence of the theory. Every way one might remove it seems ad hoc.

It is interesting because we have already discovered a quantum mechanics in the real world in which microphysical entities are represented as sets of potentials. There is a central problem about how microphysical entities that are wells of potential constitute determinate macrolevel entities. Typically, people reduce the problem to one of how to make the microlevel entities determinate after all, either by completely "collapsing" their potentiality under defined circumstances or representing them as determinate in the first place. The result here is strangely resonant with some aspects of the theory by suggesting a middle ground, that microphysical entities may remain indeterminate in some respects just as long as their indeterminacy remains compatible with the determinacy of their macrolevel context, which itself is a function of how the macrolevel individuals interact with and constrain one another. This kind of middle ground deserves investigation because it could help resolve issues with some approaches, such as decoherence approaches, to the problem.

These tight relations between possibility and actuality, along with the need to be abstract modal realists about possibility, raise questions. To explain something like the unity of the world or the direction of time, we must at least begin down an avenue of speculation about these relations between the world and its metaphysical background, no matter how few steps we take. In considering how to proceed, I am inclined, at least initially and tentatively, to follow Whitehead.

Following Whitehead, I endorse the idea that the actual world is connected to its metaphysical background by a process of becoming. The metaphysical contraction of a well of potential into a determinate and complete individual is a coming into being, a move from possibility to actuality, for a real entity. This process of becoming can be represented as a kind of vector called an *ingression* from a space of possibility to actuality. Of course, this raises the very difficult and important question of what the metric is on the space of possibilities. With Whitehead, I am inclined to think it is something like a degree of determinateness. Still, it is not an easy task to articulate what *that* is. These issues require extensive independent treatment.

For the schematic purposes of this chapter, I suggest the following definitions:

Definition 10.1: An *ingression*—A path that an individual may take from incompleteness to completeness.

Definition 10.2: A *hit*—The point on an individual's ingression at which it is complete.

Definition 10.3: An *actuality*—The states of one or more individuals when they achieve their hits.

Definition 10.4: An *actual world*—Any maximal set of interconnected hits. If *A* and *B* are individuals, then their hits interconnect when (1) *A* and *B* are at the same level of nature and they share a member individual at their hits or (2) *A*'s hit at one level is a member of *B*'s hit at the immediately higher level.

Definition 10.5: A *potentiality*—An individual *A* is *potentially X* just in case there could be an ingression for *A* in which the hit associated with that ingression is in state *X*.

Definition 10.6: A *possible world*—A maximal set of interconnected compatible combinations of individual potentials.

Definition10.7: A *possibility*—A part of a possible world.

The preceding definitions are neutral with respect to the number of actual worlds. There may only be one, our own, or it may be that our world is just one of two or more disjoint closures of interconnected hits within the space of possibilities, in which case each closure would be an actual world. Nevertheless, actuality is not a mere indexical fact about a world. An actual world is a determinate world, and to be determinate is a substantially different mode of existence than to be merely possible. As proposed in the last chapter, determinateness indicates completeness. To become fully *determinate,* an individual must have all of the slots in its receptive connection saturated by other individuals; its effective state must be determinate; and it must take a fully specific place in the receptive mesh by saturating the slots of other individuals, if such slots are available to it. *Determinateness requires full immersion in a context*. Points along an individual's ingression correspond to the degree that an individual's nomic content has been made complete (see figure 10.18).

In contrast to this, indeterminateness implies a kind of context independence. An *indeterminate* individual is one that may ingress, that is, it may exist in multiple, more definite contexts. An individual with unsaturated slots or slots saturated by individuals that leave its state indeterminate or that has an unsettled spot in the mesh is less determinate than it can be. Determinateness and indeterminateness admit of degree.

The ideas of taking on a context and of being context independent (to one degree or another) are also keys to understanding the distinction between *abstractness* and *concreteness*. To be *abstract* is to be removed from context and capable of being placed, under different determinate forms, in multiple other contexts. To be fully abstract is to be the maximally context independent part of a thing's na-

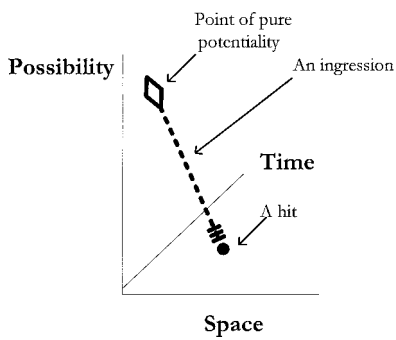


Figure 10.18 A depiction of the connection between possibility and actuality.

ture. This is why level-zero individuals, as metaphysical abstracts, can never be found in nature. Level-zero individuals are completely free of context, whereas to be part of nature is to have at least some degree of context.

In contrast, to be concrete is to find a context within a world in which individuals have causal significance. To be fully concrete is to be determinate and fully immersed in a context. To perceive a concrete thing is to become acquainted with a fully determinate thing in its full context.

Because the notion of *being immersed in a context* admits of degrees, it serves as the bridging idea that makes clear how the two poles of ideas—*abstract*, *potential*, *incomplete* and *indeterminate* on one pole and *concrete*, *actual*, *complete*, and *determinate* on the other—connect to one another and can be a matter of degree. On the far pole, a nature with no context at all is fully abstract. Because it has no contextual relations, it will be maximally incomplete. Maximal incompletion implies maximal indeterminateness because a maximally incomplete nature will not have taken constraint from any context. Therefore, an abstract thing's existence is as a potentiality that could come to exist within a number of distinct and different contexts. Furthermore, as contextual relations are incrementally specified for a nature, we can imagine it becoming by degrees less abstract, incomplete, indeterminate, and closer to actuality. At its other pole, where it takes its place within a complete context, it has become part of the actual world, having become complete and determinate and therefore concrete (see figure 10.19).

Ingressions play a dual role in this picture. On the one hand, an ingression charts a trajectory through the space of possibility for that individual into the actual world. On the other hand, we should not separate possible from actual natures, for the actual individuals are the possible individuals made fully determinate. The two poles of the ingression are not sundered, as a possible thing does not cease to be itself when it becomes actual. It merely becomes one of the many things it could have been. So an ingression also represents the complete structure of an occurrent individual's *nature* along a line of pure possibility to actuality.

I call the point at which an ingression reaches actuality a *hit*. I suggest thinking of a hit as the point of full determinateness for that ingression. If we do this, we can define the actual world as the fully interrelated collection of all hits within the larger space of possibilities. For what follows, it will be important to avoid identifying an individual's nature with the hit, which is simply the tip of its ingression.

The nature of an occurrent individual should be thought of as spread or stretched along the *length* of its ingression. Equivalently, ingressions should be thought of as schematic ways of explicitly drawing out components in the nature of an individual, where these components have various degrees of context independence. In this way, natures are taken to be complex entities containing indefinite, context-independent components, as well as definite, actual states with a complete context. The indefinite elements contain the individual's potentials. As we move up an ingression, away from the hit, we traverse an expanding well of potential and a decreasing weight of context. As we move down an ingression,

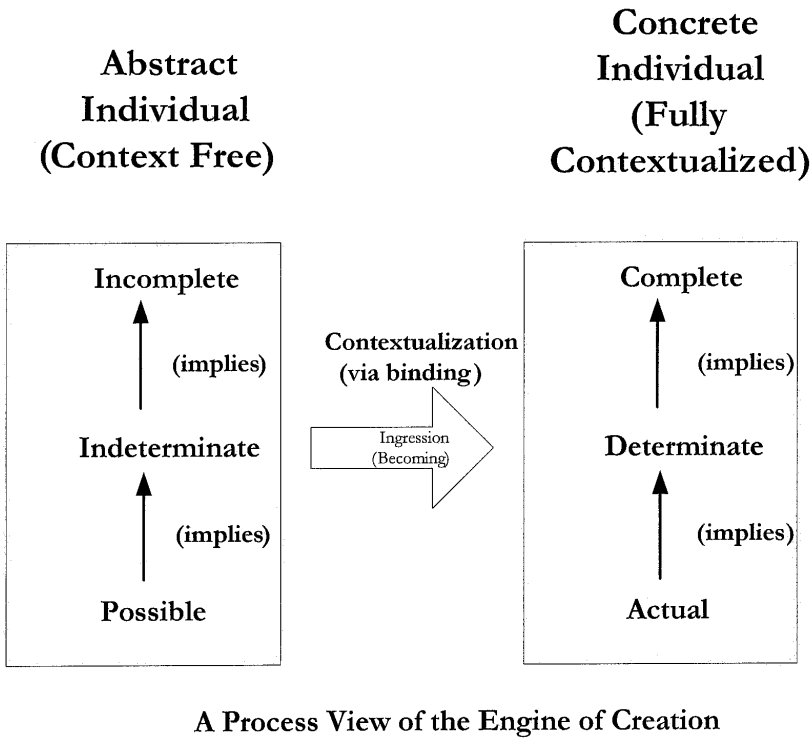


Figure 10.19 How it all fits together.

toward the hit, we traverse a shrinking well of potential and an increasing weight of context.

10.6 Space, Time, and the Unity of the World

Within a level of nature, the unity of that level comes from the overlaps between receptive slots in distinct receptive connections. The unity between higher and lower levels of nature comes from the binding relation. For an ingression to hit the world, its unsaturated slots must fill with individuals, *and* it must find a place at its own level. The unity of the world is simply the closure of the mesh across and within levels.

If that is the unity of the world, how can we understand the direction of time? We must treat space and time together. To approach this problem, I suggest that we do away with space and time as basic entities. The strategy I pursue involves reversing the common way of viewing the relations between causal connection, temporal ordering, and spatial neighbors.

Instead of assuming that the causal mesh exists in space and time, I propose that space and time are a construction out of the structure of the causal mesh.[8] In

trying to illuminate the construction of space and time, I focus on the concepts of fixed properties and asymmetric connections. These concepts are important because they can provide the building blocks of a highly regular directed graph of receptivity and constraint, and, to succeed in the construction, we need a regular pattern of receptivity from which we can reconstruct the orderly facts about space and time.

To begin, observe that one pole of an ingression corresponds to the hit, which is the point of maximal determinateness, and one pole corresponds to its origin in the space of possibility, which is the point of maximal potential. For a natural individual, in between these poles are various degrees of incompleteness: Some of the slots in its receptive connection do not contain individuals at all; and some contain individuals with indefinite states that affect the determinateness of their own states, and their potential relations to other individuals are not fully established.

The entity's ingression maps the process of becoming, but it is not a temporal becoming: It is a becoming from potentiality to actuality. To understand time, imagine that some of these ingressing individuals contain asymmetric connections, and it is a small step to speculate that the direction of time might supervene on a process of overlapping asymmetric connections. Let us call a series of individual ingressions with overlapping asymmetric connections a *cascade*. The hypothesis here is that the direction of time follows the direction of asymmetric constraint within cascades like the one depicted in figure 10.20.

To imagine the construction of time and space, imagine a world with cascades and a level-zero individual, $I_{0.2}$, in time slice 2 that is an element bound within a cascade. As a consequence of its position in the cascade, it is both asymmetrically constrained and asymmetrically constraining. Its immediate future may be organized relative to the element it is asymmetrically constraining, and its immediate past may be organized relative to the element it is asymmetrically constrained by.

In general, the other level-zero individuals in $I_{0.2}$'s cascade form a set of pivot points for organizing the spatiotemporal facts relative to it. Its past should be constructed as a function of the parts of the cascade that are fixed relative to it,

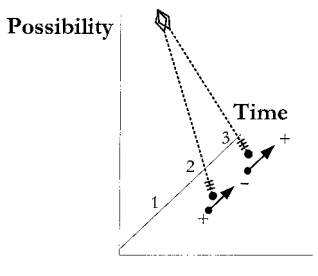


Figure 10.20 The ingression of a process.

both directly and transitively, and its future is a construction from the parts for which it is fixed. Theoretically, if the receptive structure of the world is itself regular in the right way, the facts about the cascade should naturally partition themselves into an ordered set: Each element in the cascade would correspond to a different order in a temporal history and would belong to a different time slice.

Spatial organization needs to be constructed in two broad steps. In step 1, we need to assign other individual hits to the temporal partitions constructed for the individuals in the cascade. In step 2, we need to assign distance and direction relations to the hits within each partition relative to the pivot point that defined the partition and acts as its point of reference.

As for spatial distance, the *spatial* distance from $I_{0.2}$ to other members of its own temporal partition is a way of representing the *temporal* (i.e., causal) distance between those hits and $I_{0.2}$'s cascade. Hits that are in $I_{0.2}$'s temporal partition should be placed spatially farther away or closer to $I_{0.2}$ depending on whether they are temporally farther away or closer to $I_{0.2}$'s cascade. This means that if $I_{0.2}$ is five seconds from one of its descendents in the cascade and two days from another, then a member of its temporal partition that is also five seconds from that descendent should be proportionally closer to $I_{0.2}$ than a member of its partition that is two days from its cascade. The key concept is that there is a causality condition on locality, not a locality condition on causality. Brian Cantwell Smith (1996) has expressed the intuition well. He says, "Distance is what there is no action at" (p. 190).

In more detail, we may construct a rule for determining the individuals that belong to $I_{0.2}$'s time slice using the causal relations between it, its cascade, and other individuals as follows. An individual $I_{0.j}$ belongs to the same time slice as $I_{0.2}$ just in case

1. $I_{0.j}$ is immediately bound to $I_{0.2}$ within a symmetric connection, *or*
2. $I_{0.j}$ is symmetrically bound to some *other* individual that is in $I_{0.2}$'s time slice, *or*
3. There is a descendant $I_{0.N}$ in $I_{0.2}$'s cascade that is causally equidistant from both $I_{0.2}$ and $I_{0.j}$. A descendent $I_{0.N}$ is *causally equidistant* to both $I_{0.2}$ and $I_{0.j}$ just in case there are two paths through the directed graph of asymmetric constraints such that $I_{0.j}$ is n asymmetric links from having causal significance for $I_{0.N}$ in one path, $I_{0.2}$ is itself n asymmetric links from having causal significance for $I_{0.N}$ down the other path, and there is no shorter path by which either has causal significance for $I_{0.N}$.

Clauses (1) and (2) of this rule yield a set containing all the individuals that $I_{0.2}$ is symmetrically connected to, all the individuals those individuals are symmetrically connected to, and so on. Clause (3) calculates the distance through a cascade between $I_{0.2}$ and its descendants by counting asymmetric connections. Because this causal distance constitutes the descendant's temporal distance from $I_{0.2}$, an analogous distance can be used to map its temporal distance from other individuals by counting their path of asymmetric connections to the part of the

cascade in $I_{0.2}$'s future. Essentially, clause (3) uses the equivalence between temporal and causal distance to assert that any two individuals that have the same temporal distance from $I_{0.2}$'s descendants as $I_{0.2}$ does are in the same temporal slice as $I_{0.2}$.

The rule yields a set of individuals that occupy the same time slice as $I_{0.2}$. The next task is to reconstruct the spatial neighbor relations between $I_{0.2}$ and the other individuals in its time slice using just the facts about the structure of their causal relations. To simplify the model, I assume that space is a discrete grid (in a model with continuous space, the concept of "neighbor" would need to be more technically specified). To construct spatial relations, I will try to enforce *locality* by using causality relations and projecting individuals that share a time slice with $I_{0.2}$ (and are thus directly causally relevant to one another) into space using the following three rules:

1. If individuals $I_{0.j}$ and $I_{0.k}$ share a symmetric receptive connection with one another, and there is no individual $I_{0.1}$ that shares a symmetric receptive connection with one but not the other, then $I_{0.j}$ and $I_{0.k}$ occupy the same point in space.
2. Any individuals in $I_{0.2}$'s time slice that are, like $I_{0.2}$, directly constraining $I_{0.2}$'s immediate descendent are neighbors for $I_{0.2}$ if they do not meet condition (1).
3. Otherwise, the individual is not a neighbor of $I_{0.2}$.

Rule 1 supports the possibility of pointlike entities with many properties, such as electrons whose properties of mass, charge, color, flavor, and spin are present all at a point. Rule 2 allows extension through space by establishing neighbor relations. Rule 3 prohibits assigning $I_{0.2}$ neighbors that are not immediately causally relevant to its immediate future. Starting with a given individual such as I_1, one may find its neighbors by application of rules 1 through 3. By reiterating the procedure on its neighbors, one may find their neighbors, and so on.

The most difficult part of the procedure would be assigning a specific direction to individuals within their distal rings, relative to $I_{0.2}$. Distance plus direction, together, would allow for the creation of a spatial coordinate system. The procedure would have to involve several constraints, one of which would be making sure to place each individual's own neighbors next to them.

The key concept would be that of a *signal*. A signal is a change to the effective state of an individual hit or hits within a cascade that can be propagated in some way to another cascade. It can be modeled along the lines of mark transmission as introduced by Wesley Salmon (1984). A signal requires a *signaling path*, which constitutes the individual nodes and links through the graph that the change must traverse to reach its intended target. Every potential signaling path constitutes a direction. Each neighbor of $I_{0.2}$ capable of carrying a signal can form the basis for a different direction emanating from $I_{0.2}$. Individuals who are in $I_{0.2}$'s time partition but are not neighbors of $I_{0.2}$ thus lie in a direction that is a function of the individual directions along an appropriate signaling path relative

to $I_{0.2}$. An appropriate signaling path is one from $I_{0.2}$ to some future member of the targeted individual's cascade (or, in the generalized case, a hypothetical cascade for the individual). The number of directions in $I_{0.2}$'s spatial manifold is equal to the number of neighbors it has whose cascades could continue a signaling path emanating from $I_{0.2}$.

Finally, in practice we should treat these rules as soft constraints. The "correct" spatiotemporal construction would be the one that best balances the need to preserve locality (treating direct causal significance as either a spatial sameness or neighbor relation) and the desire to obtain a simple and orderly geometry. One might give up some locality in special circumstances to make a gain of simplicity in the geometry, but ideally only a little would be acceptable.

10.7 "Fixed" Facts and the Puzzle of Asymmetric Connections

An important remaining problem is to explain how the state of an effective individual may be asymmetrically fixed for others in the way the theory requires. This is a problem for cascades, for instance. Cascades subserve the order of time, so the real problem is accounting for what it means for some facts to be fixed relative to others without assuming time. We could take the difference between symmetric and asymmetric connections as primitive, but that would be inelegant. The metaphysics would be cleaner if we could analyze all connections in a single way and derive the differences between symmetric and asymmetric ones from the analysis.

The simplest solution is to propose that asymmetric connections are symmetric connections in which individual(s) on the unconstrained end of the connection are already complete when considered independently of the nexus created by the connection. In such a case, there would be no further potentials within the individual for the other individual(s) in the connection to constrain. Under this proposal, within a nexus N containing members A and B, an individual A is *fixed* relative to B just in case A is determinate considered independently of N.

11

Is Connectivity Entailed by the Physical?

11.1 Introduction

The theory of causal significance has introduced a group of concepts important to a realist theory of causation:

- *Causal responsibility*. The idea, at least partially subjective, that one entity should be singled out as producing an event.
- *Causal significance*. The objective constraint that the existence of an entity (or the occurrence of an event) places on the possible states of the world.
- *Effective properties*. Properties that contribute to constraints on the determinate states of a causal nexus.
- *Receptive properties.* Connective properties enabling individuals to become members of causal nexii and to be sensitive to constraints on the state of nexii where they are members.
- *Binding*. A unique internal relation between two properties with incomplete natures enabling them to enter into one another's natures to become more complete.
- *A causal nexus (or nexii).* A receptive connection binding two or more determinable individuals.
- *Effective completion*. An effective determinable becoming fully determinate. Effective completion involves an effective individual binding to at least one instance of a receptive connection that is also bound to other effective individuals.
- *Receptive completion*. A receptive connection becoming fully saturated. In receptive completion, binding with an effective determinable saturates a slot[1] in a receptive connection. When a receptive connection's slots are all saturated, the connection is *fully saturated*.

- *Nomic content.* The effective and receptive properties belonging to an individual.
- *Causal laws*. Laws governing the composition of the causal nexus; that is, laws describing the compatibility, incompatibility, and requirement relationships between effective properties within a nexus.
- *Natural individual*. A primitive effective or receptive property, or a completed (i.e., fully saturated) receptive connection.
- *Levels of individuals*. The primitive effective and receptive properties form a level-zero base on top of which an open number of layers of completed receptive connections may form other individuals. Receptive connections binding level-zero individuals form level-one individuals; those binding level-one individuals form level-two individuals; and so forth.

The causal realist who accepts this conceptual framework accepts the existence of a fundamental kind of individual, the *natural individual*, which in reality is a completed receptive connection. This fundamental kind contains two causal aspects, its member effective individuals and the receptive connection binding them into the nexus. Each individual exists as a node in a fully interconnected causal mesh where it or its constituents may condition and be conditioned by other individuals.

Assuming we live in such a world, what part of it is described by physics? This chapter argues that the facts about receptive connectivity are not entailed by the physical facts and so do not ontologically supervene[2] on those facts. We have already discovered that the facts about consciousness do not ontologically supervene on the physical facts, and the arguments in this chapter constitute a first link between the theory of causation and a potential theory of consciousness.

11.2 Two Mosaics and a Mesh

Philosophers discussing causation typically discuss just two levels of pattern in the world, which I call the Humean and the nomic mosaics. The Humean mosaic is just the pattern of instantiations of the physical properties through spacetime: *This* property has *this* magnitude at coordinates S1,S2,S3,T; *that* property has *that* magnitude at coordinates S1',S2',S3',T´; and so forth. A complete description of the Humean mosaic would include the initial state of the universe, along with every subsequent state, described as property values occupying points in spacetime. The Humean mosaic assumes no objective content to the world other than some extensional facts about a pattern of property instantiations.

The nomic mosaic goes beyond the Humean mosaic by introducing modal facts into nature. The nomic mosaic adds a description of what properties *would* have been instantiated where, *had* conditions been different, bringing the world's pattern of property instantiations under reliable, counterfactual supporting laws.

The facts about the nomic mosaic do not ontologically supervene on the facts about the Humean mosaic. For example, not all circumstances covered by a general law get instantiated in the history of the universe; the gap between what the

general laws describe and what actually occurs means that a variety of physical laws, other than the ones actually true of our world, are logically compatible with the merely extensional facts about the Humean mosaic. It follows that realism about natural laws must go beyond conventionalist views of natural causation.

A causal realist's world must include the facts about the world's causal mesh and have an intrinsic causal structure that includes receptive connections between individuals. The following arguments show that the facts about the causal mesh—that is, the world's causal structure—do not ontologically supervene on the nomic mosaic, just as the facts about the nomic mosaic do not ontologically supervene on the Humean mosaic. Additionally, I argue that physics is concerned only with the nomic mosaic, and thus the causal mesh extends beyond the physical aspects of the world. In terms of the richness of their descriptions of the world, what one might call their metaphysical thickness, the Humean mosaic is thinner than the nomic mosaic is thinner than the causal mesh.

11.3 An Example: Positive and Negative Charge

As a prelude to the abstract arguments that follow, I consider a concrete example intended to clarify a common source of confusion. One kind of objection, centering on facts such as oppositely charged particles attracting one another, shows the confusion well. The fact that a positively charged proton attracts a negatively charged electron seems to be an effective fact about the charge. Yet, on the face of it, the electron seems to *receive* the attraction due to its negative charge, and we can say the same for the proton's positive charge receiving the attraction of the electron. Is this a case of physical properties being both receptive and effective?

The imagined objector simply misinterprets the situation. It is not the negative charge that receives the action of the proton. One would not properly predicate the receptiveness to these *properties* but to the individuals themselves. The electron receives the action as a whole individual, and so it is the physical object, not the physical property, that is receptive. The reception of the attraction *includes* the particle's negative charge, as part of the property complex that is the electron, but it is not received *through* the negative charge.

Correctly interpreted, the negative charge enters the story through the explanation of the particle's *response*. An electron's response is a *reaction*, not a *reception*. That an electron should receive the action of a positive charge and then have its reaction partially determined by the presence of its own charge is consistent with the claim that charges are merely effective properties. The account given here requires individuals to respond in such ways, as a nexus is holistic. Even the etymology (re*action*) suggests that charge is acting effectively on both ends.

11.4 Receptivity and Spacetime

Problems localizing receptive connections provide the first argument that the physics does not describe the causal mesh. At least during measurement, one can

designate a bounded region of spacetime in which the properties of a particle are located. It is not clear that receptive properties are strictly located in spacetime in this same way, and some good reasons exist for thinking that they are not.

In a famous set of experiments known as the EPR experiments, physicists found that the physical properties of spacelike separated particles (i.e., particles unable to send signals to one another) are correlated on measurement. These correlations are the consequence of constraints on the combinations of independently possible states for the two particles, exposing what seems like a unified two-particle system with distinct, global properties of its own. The framework introduced in the last chapter explains these correlations by appealing to the existence of a spatially distributed higher level individual created by a shared receptivity for the particles.

When it comes to such two-particle systems, the naturalness and simplicity of an explanation that proposes a shared receptivity making a higher level individual is compelling. For that very reason, it is unattractive to identify a high-level individual's receptivity with any of the physical properties. First, the instance of receptivity is shared by both particles, but instances of the physical properties are not. Also, in these experiments, the physical properties of each particle are located at their places of measurement, but the question of where to locate their shared receptivity is more problematic. Claiming that it is wholly at either of those places seems strange, as does the claim that it is solely at one of them, or even that it permeates all of the space between them. The location of instances of receptivity in spacetime is a puzzling matter, in a way that the location of physical properties in spacetime is not.

11.5 The Argument from the Possibility of Bizarre Receptive Structures

Consider a concrete dynamical system. One can argue directly that its physical description underdetermines its receptive structure. Consider the simple case of an isolated, oscillating system like that depicted in figure 11.1. It is an abstract representation of a pendulum. Physical theory's dynamical equations tell us the shape of the system's trajectory through its state space, represented by the sine-like curve in diagrams (a) through (d). The pendulum itself is a causal process consisting of many individuals receptively bound to one another at different points of time. In diagrams (a) and (b), the receptively bound individuals are those where the curves intersect with the boxes as they fall through the *XY* plane. That is, at every place that the horizontal lines of the box intersect with the curve, those four individuals share a common receptivity. The two sets of intersections yield two different receptive structures. In diagram (c), each half-circle is enveloped by an inner box representing a common receptive connection binding all the individuals within it, and the two level-one individuals thus formed are bound into a level-two individual represented by the outer box. In diagram (d), each receptive field has temporal depth, represented by the overlapping circles covering system states. These receptive fields overlap through time, one following on the other.

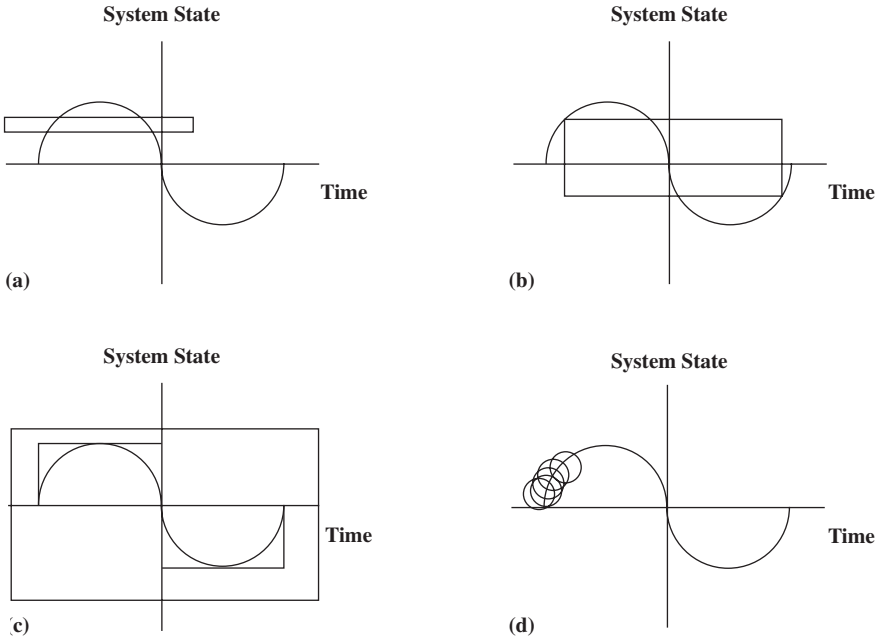


Figure 11.1 Four ways the states of a pendulum could be receptively connected.

Physical theory should be consistent with these different receptive structures because of the difference between dynamic laws and causal laws.

- *Dynamic laws* describe how the system state will change through time: They correlate system states with points in time. Physical laws are often dynamical laws.
- *Causal laws* describe the possible states of the causal nexus, which is a collection of bound states. These are the laws active within the receptive connection.

The causal laws produce the states of the causal nexus. The dynamic laws connect to the causal laws by describing patterns in the outcomes produced by the causal laws, as the pattern of receptive bindings in combination with the causal laws may place order-producing constraints on the states that individuals manifest.

These two kinds of laws, dynamical and causal, may take forms that vary arbitrarily far from each other. The rules describing the patterns in the outcomes of the causal laws may not look much like the rules describing the causal laws themselves, and baroque causal laws should exist that reproduce a nomic mosaic conforming to a dynamical function even when the receptive structures are

bizarre. If we assume, as physical theory currently does, that spacetime is a primitive entity,[3] then diagrams (a) through (d) represent different possible receptive structures consistent with what physical theory tells us.

Diagram (d) clearly represents the most intuitive structure, although even it is compatible with multiple interpretations. A structure like the one in (d) may consist of unidirectional constraint forward in time, unidirectional constraint backward in time, and polydirectional constraint.

Each diagram represents a distinct receptive structure, and nothing in physical theory will give logically conclusive reasons to assert the existence of one structure instead of one of the others. Just as pertinently, from a physical perspective the issue is one of indifference, as the dynamic equation itself does all the work in producing the experimental content of the theory. It seems to follow that physical theory is describing the nomic mosaic and does not contain the theory about the receptive structure of the world.

11.6 Physics Describes Only the Nomic Mosaic

Here I make the following argument:

1. The facts about the nomic mosaic do not determine all the facts about the causal mesh.
2. Physics describes only the nomic mosaic.
3. Therefore, the causal mesh contains facts above and beyond the physical facts.

In the following, I use the example of the *Life* world from chapter 2 to provide concrete support for premise 1. If you recall, the *Life* world is a two-dimensional world made up of cells that we can visualize as squares and that has three simple rules that fully describe its physical evolution (and hence the nomic mosaic of any particular *Life* world) (figure 11.2):

1. If a cell has exactly two *on* neighbors, it maintains its property, *on* or *off*, in the next time step.
2. If a cell has exactly three *on* neighbors, it will be *on* in the next time step.
3. Otherwise, the cell will be *off* in the next time step.

Figure 11.3 supports the premise that the nomic mosaic underdetermines the causal mesh by showing two distinct causal structures, each of which is compatible with the rules of *Life* given suitable causal laws. In essence, this is a concrete example of the argument given in the previous section of this chapter, that dynamical equations underdetermine receptive structure.

Within the figure, diagram (a) contains three levels of individuals. Each of the nine squares is a level-zero individual that may contain either an "on" or "off" value. The outer eight squares are bound together by a common receptivity that enables each of them to symmetrically constrain the others, forming a level-one

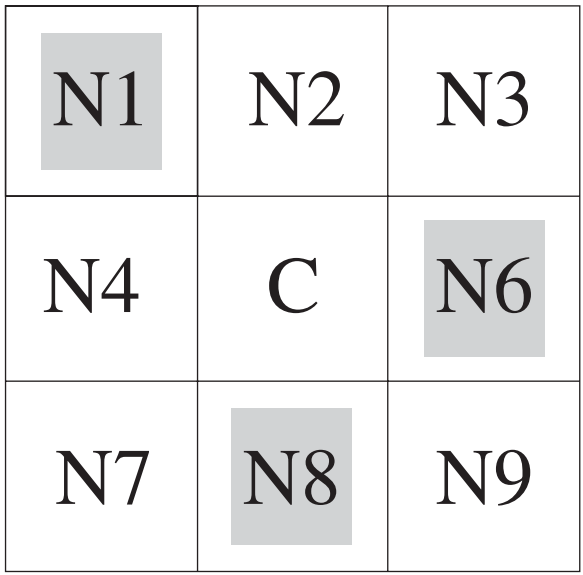


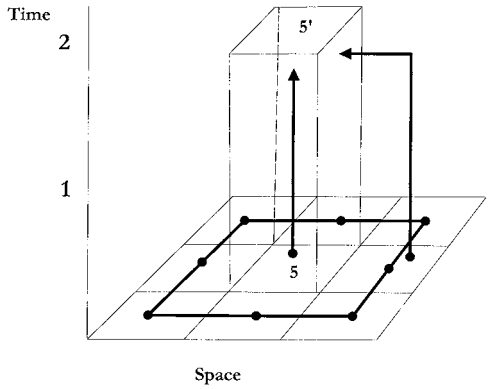
Figure 11.2 A center cell C and its surrounding neighborhood of cells. Cells N1, N6, and N8 are depicted as "on."

individual; the middle square, labeled 5, and its descendant in the next moment of time, labeled 5', are bound by an asymmetric connection from 5 to 5', and form another level-one individual. Finally, the two level-one individuals are themselves asymmetrically bound into a level-two individual.[4]

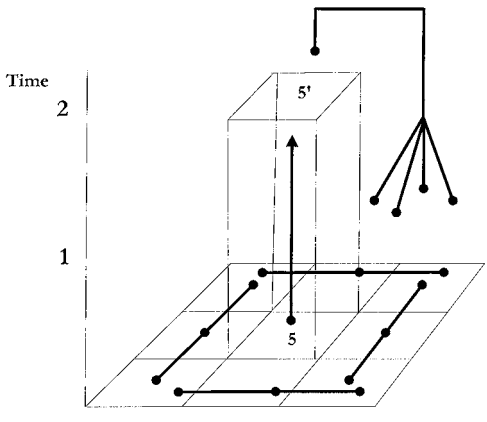
This causal mesh enables laws that straightforwardly mirror *Life*'s rules of evolution (i.e., its nomic mosaic) to produce a *Life* world. First, allow all combinations of "on" and "off" properties, without restriction, to instantiate in level-one individuals. A *Life* world can arise from three laws that govern how these level-one individuals interact within level-two individuals. To express these laws, use a function *v(x)* that accepts an individual as an argument, returns the number of "on" properties it contains, and uses an operator set (*x, y*) that accepts an individual *x* and sets the number of "on" properties bound inside it to *y*. Finally, call the two-place level-one individual connecting cells 5 and 5' $I_{2.1}$, and name the level-zero individual that is cell 5 $I_{0.5}$. The three rules of *Life* translate into the following three laws:

1. If $v(x) = 2$, then set $(I_{1.5}, 2*v(I_{0.5}))$
2. If $v(x) = 3$, then set $(I_{1.5}, v(I_{0.5}) + 1)$
3. Otherwise, set $(I_{1.5}, v(I_{0.5}))$

Law 1 corresponds directly to the two-neighbor rule, law 2 corresponds to the three-neighbor rule, and law 3 corresponds to the "otherwise" condition in *Life*.



(a) A simple receptive structure whose causal laws correspond directly to Life's physics



(b) A more complicated receptive structure with more gerrymandered causal laws

Figure 11.3 Diagrams (a) and (b) represent two distinctive receptive structures for a *Life* world.

As well as producing lawlike patterns of *effective* property instantiations, this example also illustrates how receptive connections in a realistically orderly world will also instantiate in a regular, lawlike way.

Diagram (b) in figure 11.3 represents a different causal structure that can also produce a *Life* world under suitable causal laws. In diagram (b), the outer eight cells are broken up into four overlapping level-one individuals, each with a three-place receptivity. Also, each of these level-one individuals, along with the level-one individual constituted by the asymmetric binding of cells 5 and 5', is bound into a level-two individual through a five-place receptive connection. The important differences between diagrams (a) and (b) are at levels one and two, where

the receptive structure of the diagram (b) world is different from the diagram (a) world. Consequently, the laws required to make this universe produce a *Life* world are more complicated. I do not list them all, but, for example, the three-neighbor rule of *Life* translates into a series of rules such as:

If $(v(I_{1.1}) + v(I_{1.2}) + v(I_{1.3}) + v(I_{1.4}) = 3)$ &
$((v(I_{0.1}) = 0)$ &
$(v(I_{0.3}) = 0)$ &
$(v(I_{0.7}) = 0)$ &
$(v(I_{0.9})) = 0)$, then set $(I_{1.5}, v(I_{0.5}) + 1)$

If $(v(I_{1.1}) + v(I_{1.2}) + v(I_{1.3}) + v(I_{1.4}) = 4)$ &
$((v(I_{0.1}) = 1)$ &
$(v(I_{0.3}) = 0)$ &
$(v(I_{0.7}) = 0)$ &
$(v(I_{0.9})) = 0)$, then set $(I_{1.5}, v(I_{0.5}) + 1)$

If $(v(I_{1.1}) + v(I_{1.2}) + v(I_{1.3}) + v(I_{1.4}) = 4)$ &
$((v(I_{0.1}) = 0)$ &
$(v(I_{0.3}) = 1)$ &
$(v(I_{0.7}) = 0)$ &
$(v(I_{0.9})) = 0)$, then set $(I_{1.5}, v(I_{0.5}) + 1)$

If $(v(I_{1.1}) + v(I_{1.2}) + v(I_{1.3}) + v(I_{1.4}) = 4)$ &
$((v(I_{0.1}) = 0)$ &
$(v(I_{0.3}) = 0)$ &
$(v(I_{0.7}) = 1)$ &
$(v(I_{0.9})) = 0)$, then set $(I_{1.5}, v(I_{0.5}) + 1)$

If $(v(I_{1.1}) + v(I_{1.2}) + v(I_{1.3}) + v(I_{1.4}) = 4)$ &
$((v(I_{0.1}) = 0)$ &
$(v(I_{0.3}) = 0)$ &
$(v(I_{0.7}) = 0)$ &
$(v(I_{0.9})) = 1)$, then set $(I_{1.5}, v(I_{0.5}) + 1)$

And so on. The causal laws must take on this much more complicated form to reflect the true causal structure of the world. For example, the conjunctive conditions are needed to reflect the fact that there are overlaps in the three-place individuals that constitute a cell's neighborhood.

This difference in laws between the two diagrams is required because different individuals with different effective states exist in the two worlds. Basically, the world of diagram (a) simply has no laws to handle the individuals of diagram (b), and vice versa. There's no fact of the matter about what constraints each world's individuals would present in the other world, and so no fact of the matter about

what kind of effective properties those individuals would realize in the other world. The two worlds literally instantiate distinct effective properties and receptive facts even though they both result in *Life* worlds.

One might object that these two worlds are the same world, appealing to their nomic equivalence. That objection ignores the realist character of the theory of causal significance. The two diagrams contain, quite literally, different receptive connections and so different kinds of individuals. Therefore, the causal laws describing the causal structure of those two worlds *quantify* over different individuals. To ignore this difference in the structure of individuals and in their connections to each other is to slip back into a Humean irrealism about the world's causal structure. As we saw in chapter 8, Humeanism creates radical metaphysical and epistemological problems and is worth avoiding.

The two diagrams represent different causal structures that instantiate an identical nomic mosaic. The take-home point is direct and powerful: The causal structure of the world is not determined by the nomic mosaic it instantiates, where we may understand its nomic mosaic as the simplest description of the lawful ways its effective properties instantiate through spacetime. In fact, it is the other way around. If we are being realists about causation, the causal structure of the world will include the patterns of receptive connection between individuals, and different causal ontologies may produce an identical nomic mosaic as long as their causal laws vary appropriately.

Precisely stated, a world's causal facts and laws are not entailed by, and therefore *do not ontologically supervene* on, the set of facts including just (1) instantiations of lower level effective properties, specified as merely distinct (and, one could add, varying in quantity), and (2) the laws that describe the regularities and correlations between their instantiations through space and time. In the following, I use this conclusion as an important part of my argument that the facts about the receptive properties of our world are, at best, only suggested by physical theory. The complete premise set for the argument is:

1. Physics *at least* designates the low-level effective properties (chapter 9, section 7).
2. When proposing the physical laws, we *at least* specify regularities in the instantiation of effective properties (chapter 9, section 7).
3. Physical theories represent the least set of properties that we need to explain experimental results (simplicity constraint).
4. The receptive structure of the world does *not* ontologically supervene on the facts about the pattern of instantiations of the low-level effective properties, designated as merely distinct (the previous argument using the *Life* world).

Imagine now that we have a theory P that meets the requirements of premises (1) and (2). This means that it describes the lawlike ways that effective properties instantiate, but nothing else. P is, in fact, the theory of the nomic mosaic, and, by (1) and (2), we know that physical theory at least encompasses P. Now,

let P^* designate an extension of P that includes the causal laws involving the receptive structure of the world. By premise (4), we know that P^* must be a *proper* extension of P. The simplicity condition, premise (3), means that physical theory encompasses P^* only if the experimental content of P^* is greater than that of P.

To argue that the facts about the causal mesh go beyond physical theory, show that the experimental content of P is equal to that of P^*. Let E be an experimental design used to measure some physical property, *pp*. The experimental design E is adequate just in case the experimentalist can predict how his measuring devices will vary with the presence of *pp*. Recall that the important aspects of our measuring devices are those that make an effective difference to perception. The scientist's goal is to use the structure of the experimental situation E to derive conditionals of the form, "If property *pp* is present in these circumstances, then the measuring device will be in effective state *s*."

The problem reduces to one of establishing the requisite subconditionals governing significant links in the magnifying chain. By hypothesis, P alone will be sufficient to derive the character of the subconditionals, as it describes the covariations between effective states at every stage. It follows that P is sufficient to derive the experimental content of our physical theories. Given simplicity (3), physical theory should not include the additional content which is in P^*, being the laws governing the structure and behavior of the causal mesh and, particularly, receptive connections and causal laws. Such facts would complicate physical theory while being explanatorily superfluous to its guiding theoretical concerns. Physical theory should fail to reveal the world's receptive structure for basically the same reasons *Life* rules fail to reveal a receptive structure for *Life*.

11.6 From Receptive Connections to Carriers

The inability of facts about the causal mesh to increase the empirical content of *physical* theories in no way implies that receptive connections, natural individuals, or the causal laws are causally irrelevant. How could they be? Instead, these arguments imply that a causal realist's world contains fundamental causal facts lying beyond the standard concerns of physical science. Additionally, we should not assume that its failure to increase the empirical content of physical theories implies an inability to increase the empirical content of science in general. In a Liberal Naturalist world, there will be scientific theories whose empirical content includes nonphysical facts about experience, and the theory of causal significance has a role to play in predicting some of those nonphysical empirical facts.

The payoff from this chapter is that a Liberal Naturalist can now place the physical facts within a larger causal context. Physics, by concerning itself with a description of the nomic mosaic in the simplest possible terms, leaves the world's receptive structure, its connective tissue, so to speak, as something we understand only implicitly through the way we *use* physical theory. In particular, because the world's receptive structure is only suggested in physical theories, a variety of structures for the underlying causal mesh are compatible with those theories. It

seems, then, that physical theory is explicitly about only an aspect of the causal mesh, its effective side.

The full ontological structure of the world is a mesh of individual causal nexii, each possessing nomic content constituted by receptive and effective aspects within the nexus. The differences between what physics describes and what the world contains opens avenues for developing Liberal Naturalism. This is a great step forward, but the Liberal Naturalist still needs to take another, equally important, one. We must squarely face the force of the arguments in chapter 2 and honestly answer their challenge. The next chapter takes this step by introducing the *Carrier Theory of Causation*, a theory that will bring experience back into the picture.

12

The Carrier Theory of Causation

> *Even if there is only one possible unified theory, it is just a set of rules and equations. What is it that breathes fire into the equations and makes a universe for them to describe?*
>
> Stephen Hawking, *A Brief History of Time*

12.1 Circularity in the Causal Mesh

Chapter 2 presented the antiphysicalist arguments through an analogy to *Life*, an artificial world. In it, I explained why the "on" and "off" properties at the heart of *Life's* physics, its schematically characterized categories, were unable to support consciousness. Consider the question, *What is it to be an "on" property within* Life? We only need to cite its distinctness from the "off" property and that its patterns of instantiation exhibit the lawlike regularity prescribed by the rules. This schematic account entirely encompasses the categorical being of an "on" property within *Life*.

The last few chapters presented ideas that enrich the minimal view of causal content explicit in the bare physics of *Life*. I have argued for effective and receptive faces to causation, have given a positive account of the relations between them, and detailed the distinct contributions they make to the causal character of a world. Simplicity dictated the form of the Theory of Causal Significance, and it proposed only that nature's deep structure consisted of receptive connections, effective properties, and a metaphysical background of possibility. Even space and time evaporated as fundamental entities. Still, with respect to consciousness, the new metaphysics remains limited by the fundamental shortcoming of the *Life* world because we still have a merely schematic understanding of what it is to be a world with causal content.

To describe effective properties, we need to (1) make names for the determinable types; (2) stipulate that the names designate distinct entities; (3) define a range of determinate values for each determinable; and (4) define how the presence of each effective property within a shared receptive connection contributes to the constraints on the overall state of a causal nexus. Descriptions meeting these conditions create identity conditions for each property and for the system of properties as a whole.

I call the kind of circularity involved *contrastive circularity*. A contrastive circularity exists when we stipulate a set of distinct elements and we use the stipulated distinctness to create a relational story yielding the rest of each element's identity conditions. For any kind of element *X*, we may answer the question "What is it to be an *X*?" by citing its distinctness from the other elements of the system and a (perhaps fuzzy) set of critical external relations to other elements within that system.

Contrastive circularity is enough to distinguish effective properties from one another, such as "on" from "off," but a circularity of a different kind helps to distinguish receptive and effective properties. Receptive and effective properties together form a causal nexus. Here are the relevant definitions again:

- A *causal nexus* (or *nexii*), A receptive connection binding two or more determinable individuals.
- *Effective properties*. Properties that contribute to constraints on the determinate states of a causal nexus.
- *Receptive properties.* Connective properties enabling individuals to become members of causal nexii and to be sensitive to constraints on the state of nexii where they are members.

The binding relation mentioned in the definition of *causal nexus* is an internal relation through which the effective and receptive properties may complete one another's inherently incomplete natures. Receptive and effective properties, therefore, do not relate categorically in a merely negative, contrastive way. Receptivity positively presupposes effectiveness and vice versa. My name for this kind of circularity is *compositional circularity*. A circular relation is compositional when each element is partially defined by the way that it presupposes the other as a positive component in its own nature (see figure 12.1).

Of the two kinds of circularity, contrastive circularity in particular is a symptom of schematic thinking. Reflection on circular contrasts naturally raises questions about how these categories come to be in a world. After all, if *A* is defined as that which is distinct from *B*, *B* is defined as that which is distinct from *A*, and *no other* categorical facts are true of them other than relations they share, it seems that the existence of each presupposes the existence of the other. The circularity has the logical feel of a vicious regress, and questions about how such properties get their footholds on existence in the first place forcefully present themselves.

Simplicity is not the only theoretical virtue. In my construction of the Carrier Theory of Causation, I am as concerned with matters of intelligibility and uniformity as I am with simplicity. By *intelligibility* I mean the construction of a model for the circular schema that is somehow based on things within our experience of the world or things that can be conceived in analogy with our experience of the world. By *uniformity with our knowledge,* I mean basing the model on the evidence of how similar problems are solved in other domains. Understanding the ontological ground for the existence of these kinds of circularly in-

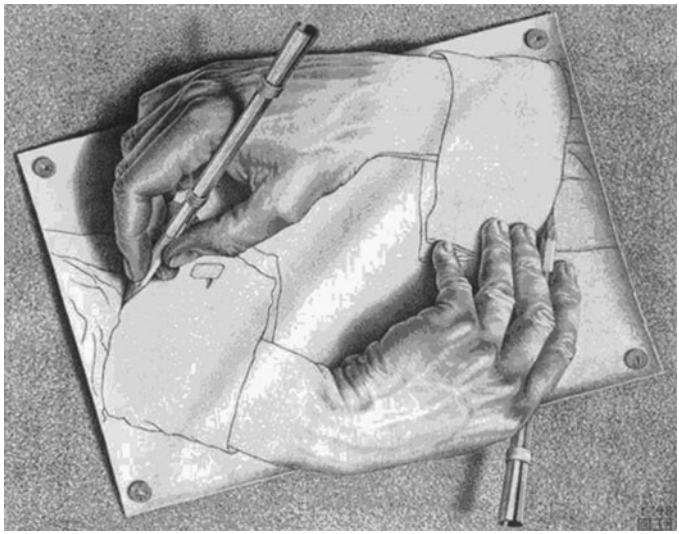


Figure 12.1 M. C. Escher's "Drawing Hands" © 2004 The M. C. Escher Company, Baarn-Holland. All rights reserved.

terdependent properties in conformance with simplicity, intelligibility, and uniformity is the chief task of this chapter.

12.2 Circularity Instantiated

Although puzzling, circularity is not deadly. The world is full of systems that realize circularly interlocking categories. The *Life* world itself provides an example, as people implement finite *Life* worlds in a variety of ways. The original *Life* world was implemented on a checkerboard using checker pieces, and the implementations most familiar to people today are on computers. I think we can gain some insight by examining how a checkerboard can become a *Life* world.

Life worlds are parasitic on the presence of properties whose categorical natures[1] are *not* wholly defined by the *Life* schema but which already embody the distinctiveness needed to instantiate *Life* properties. For example, in a checkerboard implementation of *Life,* the red and black checkers play the role of the "on" and "off" properties. The color distinction between them, a distinction that is not intrinsic to the conceptual schema that defines *Life*, is an extrinsic carrier of the needed distinctness between the "on" and "off" properties. The checkerboard carries the geometry of *Life's* two-dimensional space and allows us to create coordinate identities for the places in the world using the squares on the checkerboard. Finally, the intentions of the human beings manipulating the board guarantee the lawfully reliable patterns of instantiation between the properties (to the extent that human intentions are lawfully reliable). Once again, these human intentions have a nature that outruns the terms of the *Life* schema.

The form of the solution seems to be this: The circular properties of *Life* exist because we find elements with preexisting kinds of properties that, when put in the right combination, carry the relations the *Life* schema dictates. I call these elements *carriers*.

> Definition 12.1: *Carriers*—Objects or properties whose natures outrun the categories of a given schema but which can enter into the appropriate relations with one another when put into the proper combinations.

Carriers are also present in computational implementations of *Life*. In those cases, the physical distinctness of the machine states carries the distinctness between the "on" and "off" properties. The major difference between the computational case and the checkerboard case is that the tokens in the computational case can directly carry the lawful reliability of the "on" and "off" properties, bypassing reliance on human intentions.

One naturally sees the carriers as having natures that are at least partially extrinsic to the *Life* world they are implementing. For instance, the specific distinction between the redness and blackness of the checkers goes beyond the requirement of generic distinctness for "on" and "off" properties. This observation applies to more than just the facts about their distinctness. The causal contents of the elements of the system transcend the causal powers attributable to them as "on" or "off" properties. This suggests a useful definition of what it is to be an extrinsic property, relative to a system the property occurs within.

> Definition 12.2: *P* is an extrinsic property *within* a system *S* if, and only if, *P* is instantiated within an instance of *S* and *P* has a nature that is not exhausted by its relations to other elements as they are defined within *S*.

The idea of being an *extrinsic* property *within* the system *S* is different from the idea of being a property that is *intrinsic to* the system *S*. The properties intrinsic to a system are those whose identity conditions are grounded in relations between an object (or objects) in the system and other elements of the system.

Thus properties *intrinsic to* systems both constitute and presuppose the existence of the systems they exist within, which makes them unlike properties that are *extrinsic within* those systems. Stating the idea in another way, the properties intrinsic to a system have no nature entering into their categorical being except those defined by a system of the relevant type. In this sense, the categories of their systematic contexts exhaust their categorical natures.

For example, the properties of being "on" and "off" are properties intrinsic *to* the system of concepts defining *Life* because such properties both constitute and presuppose the existence of a *Life* world. In contrast, being a red checker is an extrinsic property *within* some implementations of the *Life* world. Redness, although present within some implementations of such worlds, is neither dependent on nor definable by the system of concepts constitutive of *Life*. The categorical being of redness, unlike the being of "offness," is independent of the conceptual scheme demarcating *Life*.

Finally, a property being extrinsic within a system does not entail that it is extrinsic *tout court*, as it may be intrinsic to some *other* system. This is the case with computational states that may implement *Life*, as those states, qua computational states, are extrinsic within *Life* but intrinsic to the computational system they exist within.

Contrasts. One particularly important feature of the examples is how extrinsic properties may implement the stipulative contrasts within the system in virtue of what I call internal contrasts.

> Definition 12.3: An *internal contrast* exists between *A* and *B* if, and only if, there is a comparative relation *R* such that necessarily, if *A* exists and *B* exists, then *R*(*A*,*B*).

Any two intrinsically distinct things have at least one internal contrast, the relation of being distinct. Additionally, internal contrasts are a kind of supercategory here. They include the *stipulative contrasts* observed to exist within schemas and can include other kinds of contrasts, too. For example, rather than composing the involved properties the way that stipulative contrasts do, internal contrasts instead may be *consequences* of the natures had by the properties. In fact, internal contrasts that are not themselves stipulative are good candidates to *implement* systems of stipulative contrasts: As long as both *A* and *B* are present within a world, the relation *R* that holds between them may *carry* a stipulative contrast. Thus, even though the entities composed by stipulative contrasts presuppose one another, they may come into existence within a system all at once by being a consequence of the nonstipulative internal contrasts of the carriers.

Consider red and black checkers again. The redness and blackness constitute an *internal* contrast able to carry the *stipulative* contrast between "on" and "off" properties in a *Life* world. Although internal, the contrast between the colors is not stipulative because we cannot reduce the natures of the phenomenal colors to a structure of pure difference relations holding between them. Finally, it is worth anticipating that *R* may be a more complex kind of relation than mere distinctness, although I have not surveyed any such examples yet.

Reflection on the variety of circular systems that actually exist in the world strongly suggests that each exists in virtue of carriers that are extrinsic to the system. For example, a chess game consists of a circularly interdependent set of types: pawns, rooks, kings, queens, and so forth. Each type is defined by the set of allowable moves it may make within the game as a whole. Without the context of the game, no particular type could exist. The circularity between their categorical natures makes it look as if the existence of each part of the game presupposes the existence of the game as a whole, which, in turn, presupposes the existence of the parts (Sellars 1963b; Haugeland 1993).

Why isn't the circularity of chess categories deadly? The reason chess games can actually exist is that each implementation takes advantage of external properties to introduce, piecemeal, the distinctions and dependencies that are defined

whole cloth within the conceptual system. For instance, pieces that are recognizably physically distinct are used to stand in for tokens of distinct types. Board positions are defined relative to a physical space within which the board has geometric relations to human players. These extrinsic distinctions and relations allow players to form intentions to move the pieces only in accord with the rules governing the types of which they will be tokens. When each of these extrinsic factors slips into place in a way that carries the circular relations of the system, a chess game exists.

Notice how the existence of a chess game seems *essentially* dependent on these extrinsic properties and relations: Were it not for extrinsic factors that have internal relations able to carry the circular identity conditions, the game simply could not exist. If the pieces were not physically distinct, the players would soon lose track of which piece was standing in for each type, and they could not form the proper intentions to play them according to chess rules. If the board did not exist in physical spatial relations to the players, the pieces could not be set up in accord with the rules, and the players could not decide the legality of moves. In other implementations, such as computer implementations, some other set of extrinsic properties always performs the carrier role.

Many more examples exist. A computer's logical components are circular because functional relations between the elements of a system are what define computational elements. As before, computer programs may exist by being carried by physical states that have natures extrinsic within the computational system. As we widen our view to other conceptual systems—economics, biology, and psychology are examples—we see the same pattern repeated.

In economics, what things count as goods and services? To a first approximation, goods and services are those things that consumers and producers barter. Who are the consumers and producers? Consumers and producers, in their turn, are people occupying distinct positions in the system of bartering for the goods and services.

In biology, organisms pass heritable characteristics through their genes. A heritable characteristic is one that parents pass from their generation to later generations. A parent, in its turn, is an organism that passes along its genes, or a significant portion of its genes, to the young.

In psychology, beliefs, desires, and perceptions are at least partially definable in terms of their functional role within the cognitive economy. An entity's functional role is its disposition to interact with other entities in the system.

In each case, a closed or semiclosed system of theoretical concepts exists, many of which are directly or indirectly circularly dependent on one another. This circularity is hardly fatal or even objectionable. The reason the circularity is innocuous is that, in each case, extrinsic properties exist within the systems, and these properties have internal contrasts between them. These contrasts help carry the circular dependencies. In economics, we can find extrinsic carriers by appealing to the desires, needs, and opportunities of individuals in the wider social system of which the economy is part. In biology, we can appeal to the mechan-

ics of molecular biochemistry. In psychology, we can appeal to computational or dynamical properties of neural systems and the way these properties help the organism survive in an ecological niche.

12.3 The Circularity of Physics

Reflections on examples such as these lead one inevitably to concepts with wider and wider spheres of application. In the case of the natural sciences, this expanding arena of circularly looping systems traces the same path as intuitive expectations of reduction. When we look at a circular system of concepts, we find that its instances are carried by objects with properties extrinsic within that system but intrinsic to some other system. Inevitably, these other systems are themselves circular, partially or completely, and thus we find them carried by yet another set of objects with properties extrinsic within them. From economics, we look to social relations of a broader sort, then from those to psychology, ecology, and biology, then to chemistry, and finally to physics.

If this way of looking at things is correct, these higher level domains are not just *in fact* realized by the entities of some other domain, the domain of the carriers, but they *need* carriers to get their foothold on concreteness. The existence of carriers is an essential ingredient, a metaphysical presupposition, for the satisfaction of circularly interdependent systems of categories.

When we reach physics, we find the same kind of circularity as in other, less fundamental, sciences, and the pivotal, required role for carriers raises questions. We can easily see the circularity in physics by asking questions about the identity conditions on the basic physical entities. These conditions are broadly functional. What it is to be a photon, for instance, is to play the functional role in our environment that photons play in physics. What it is to be charge, mass, or spin is to be distinct from the other physical properties and to nomically instantiate the pattern of regularities prescribed by the laws (again, in our environment). What it is to be gravity is to play the role gravity plays, and similarly for the other basic physical properties. As a result, physics incorporates circularity, just as all functional systems do.[2]

The circularity of physical concepts leads to the question, *What is extrinsic within physics?* That is, what carries the contrasts and relations needed to satisfy our system of physical concepts?[3] Taking a hard line here, insisting that nothing carries the physics, is unprecedented and problematic. It is unprecedented in that the extrinsic properties in other circular systems are not spandrels but elements required for the instantiation of those systems. It is problematic in that the resulting metaphysics seems unintelligible, if looked at too closely. The metaphysics requires a system of contrast, and relations between the contrasts, in which these contrasts have no carriers. Without carriers, it requires a notion of pure contrast, contrasts that seem not to be contrasts *between* anything. The idea seems to melt away before the mind's eye, like an echo issuing from no originating voice. The champion of such a metaphysic takes on a large unmet burden

in trying to explain it. Unfortunately, the sophistication of the circle in physics makes it easy to overlook the problem, or feel it less vividly, and current philosophy tends not to press the issue. I think the *Life* world makes the strangeness clearer (it is, after all, just a toy physics). Enlarging the circle makes the carrier problem easier to ignore but does not make it go away.

Taking the different tack of answering that physics has no ground level at all—it's turtles all the way down—is conceptually and logically a little less problematic. There perhaps *could* be such a world. The problem with the suggestion as an account of our world is empirical: Planck's constant seems to put a limit on how finely space, time, and matter can be divided. Below that level, there is no sense postulating further physical structure. Our world seems to have a fundamental physical floor.

If there is a physical floor, standards of intelligibility demand that there must be a set of properties that are extrinsic within physics. To find these properties, it will not help to appeal to some wider *system* of properties or to circle back around in a constructivist way to society or to human psychology. Those maneuvers just enlarge the circle, presenting the same problem again. A proper solution will be one that short-circuits the puzzle, not one that moves it to a new arena. What the world needs from a carrier of physics are properties whose being would be extrinsic within *every* such system and yet which still have the requisite internal relations to one another. For physics, we need *ultimate carriers*.

The properties answering to this description are best thought of as properties that are intrinsic *tout court*. A property whose categorical nature is extrinsic within every *system* of properties is simply one whose being is intrinsic at least partly to *itself* rather than to its contextual relationships. That is, it is a property that we cannot understand in purely systematic terms without leaving something out. The least strained way of understanding the physics of the world is to suppose that some kind of intrinsic property carries each effective property, where we understand intrinsic as intrinsic *tout court*, rather than intrinsic to a system.[4] Having come to this end, perhaps it will help to summarize the kinds of properties I have discussed so far:

1. Property intrinsic *to* a system: A property whose identity conditions are given entirely by relations to other entities within some system to which it belongs (e.g., the "on-ness" of a *Life* cell).
2. Property that is intrinsic *tout court:* A property that is not intrinsic to any system (e.g., phenomenal redness).
3. Property extrinsic *within* a system: A property that is present within an instance of a system and that has a nature not exhausted by its relations to other elements as they are defined within that system (e.g., the redness of a checker used to instantiate the "on" property within a game of *Life*).

The carriers of physics will be intrinsic *tout court* and so extrinsic within the world as it is defined by physics. Additionally, to act as carriers of the effective properties described by physics, these intrinsic properties must have internal con-

trasts with one another that mirror the features and relations of physical properties: patterns of distinctness *within* determinable families, patterns of distinctness *between* determinable families, variations in magnitude, and relations of compatibility, incompatibility, and requirement.

How many carrier candidates can there be? The phenomenal qualities of phenomenal consciousness are perfect candidates.

1. Phenomenal qualities are intrinsic *tout court:* One cannot understand what it is to be phenomenal yellow in terms of a system of relations (that is one of the lessons of the antiphysicalist arguments). Their intrinsicness is plausibly what makes qualia the funny things that they are and what makes full knowledge of them attainable only by acquaintance with them. Formally, their natures are intrinsic in the sense that a phenomenal property is not categorically constituted by the structure of relations into which it enters.
2. The failure of phenomenal properties to ontologically supervene on the physical while still being part of the natural world means that they can plausibly meet the condition of being extrinsic within the physical world.
3. Phenomenal qualities also plausibly support the required kinds of nonstipulative internal contrasts. The differences between phenomenal qualities are grounded in the differences in their intrinsic natures so that, necessarily, if they exist, then the differences obtain. For example, distinct sounds exist such that, if each exists, then they are necessarily distinct types of sound. It seems like a trivial point, but it is very important.
4. Phenomenal properties fall into natural determinable families such as colors and sounds, with intrinsic patterns of distinctness within and between families.

The internal contrasts between phenomenal properties are very important because phenomenal properties enter into much more complex internal relations than mere difference or distinctness. Of special importance is that they can possess internal *scalar* relations. Scalar comparisons within (but not necessarily between) phenomenal groups such as colors, sounds, tastes, and so forth come naturally to us. For example, some sounds are louder than others, some colors brighter than others, and some tastes are more sour than others. The most natural way to think of these groups is in terms of phenomenal spaces that they instantiate, with natural orderings of various types between the elements of these spaces along an intensity metric, such as brightness or loudness, internal to the kind of property.

The reality of scalar relationships between familiar phenomenal properties suggests that some other kinds of phenomenal properties, if they existed, could carry the kinds of quantitative variations required by physics. With this in mind, the Liberal Naturalist proposal would be that there are alien phenomenal properties in which an internal contrast between phenomenal quality A and phenomenal quality B exists such that, when they both exist, necessarily $A > B$ is true along

some natural metric. Properties such as *A* and *B* (and presumably other members of the phenomenal group they belong to) may carry the more complex kinds of quantitative contrasts required by physics.

Continuing this train of thought, a variety of compatibility and incompatibility relationships hold between phenomenal properties and possible phenomenal fields. A straightforward case is the postulated red/green incompatibility in our color space.[5] Much more subtle and sophisticated kinds of compatibility restrictions also show in experience, restrictions that apply to whole fields or subfields of a phenomenal manifold. For instance, it is not clear that one could simultaneously experience the Necker cube (shown in figure 12.2) as having face up and face down in the same visual manifold. If this restriction holds, then it is a very interesting kind of exclusion relation, one that incorporates the semantics of the conceptualization right inside the formation conditions on the qualitative experience.

The physical explanation of these incompatibilities in terms of opponent processes in the brain does not undermine or compete with the hypothesis that the phenomenal properties have these intrinsic relations. If phenomenal properties carry effective properties, then it is ultimately these intrinsic relations between the phenomenal properties that form the basis for the opponent behavior described by physical theory. The physical theory is a reconstruction of the results of the carrier's causal behavior from an external and structural point of view. It is as if the natural individuals were objects thrown at one side of a curtain, with us on the other side, and against which we can only place our hands and feel the impacts. Our physical descriptions explain the patterns of indentation in the curtain by supposing the objects to have certain shapes and compatibilities, but we are blind to the substantial nature of the objects.

Also, certain phenomenal properties might necessitate other phenomenal properties. For example, colorless instantiations of shape might be impossible, so the existence of a shape property might necessitate the existence of a color. On an even finer grained level, one might postulate that the existence of a hue necessitates the existence of a brightness value (no hue without brightness). In the gap

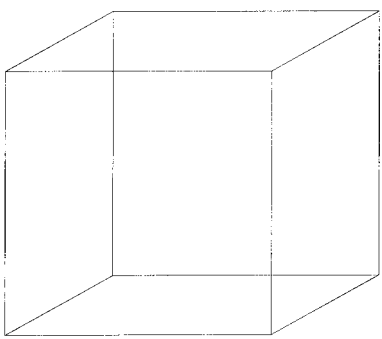


Figure 12.2 The Necker Cube. Notice how the two perceptual interpretations of its orientation seem to exclude one another.

between simple red/green incompatibility and the very subtle Necker cube face up/face down incompatibility, there might be a whole host of subtle and interesting relations of exclusion, compatibility, and necessitation between possible kinds of phenomenal properties. These kinds of relationships would be needed by a carrier that could drive the effective side of causation, as it is these relationships that would carry the natural possibilities, impossibilities, and necessities of physics.

A note of caution: The Liberal Naturalist claim would *not* be that the phenomenal qualities of human consciousness exist at the microphysical level, carrying the effective dispositions of microphysical entities. The Liberal Naturalist hypothesis here would be like the panexperientialist hypothesis: that alien phenomenal intrinsic properties exist, properties in an abstract sense like the qualia of our own consciousness, that carry the effective dispositions of the world's basic natural individuals. Perhaps these phenomenal properties are to our experienced phenomenal properties as brightness, hue, and saturation are to a full-fledged color. Perhaps the qualities at each level are irreducible. In no case are qualities of the mind purported to exist in nonmental contexts.

The abstract sense that the alien qualities would be like the qualities of our consciousness would come to precisely this: They would be intrinsic *tout court*; they would be determinables and belong to families of determinables; they would share both simple and sophisticated internal contrasts with one another; and they would have intrinsic relations of compatibility, incompatibility, and requirement.

The ways they would be different from the qualities of our consciousness would be these: Their specific characters presumably would be entirely different from those of our own qualia; the internal contrasts that hold between them would organize them into very different kinds of phenomenal subspaces; and they presumably would not be appropriate vehicles for representation and thought.[6]

But there remains one last foreboding question about the similarity of these proposed phenomenal qualities to the qualities of our own conscious experience. Would these alien qualities be *experiential*, like the qualities of our own consciousness?

12.4 The Experiencing of Phenomenal Individuals

The physical properties are the effective properties, so by proposing phenomenal carriers for the physical properties, we would account for one-half of the nomic content possessed by natural individuals. The other half of their nomic content is the irreducible receptivity in their nature, which binds effective individuals, thereby creating causal nexii. What carries receptivity?

Physics suppresses the receptivity of the world in its theorizing and thereby leaves out its receptive structure (chapter 11). The addition of receptivity to the effectiveness of physics brings a compositional circularity into the causal character of the world, magnifying the problems that arise merely from the circular contrasts of the effective properties alone. An individual's nomic content as a whole,

not just the effective aspect of it, needs to be carried, so receptivity needs a carrier as well. My fundamental proposal is that receptive properties are carried by inherently experiential properties: *Experiencing itself carries receptivity*. This is the central thesis of the Carrier Theory of Causation:

> *The Central Thesis*: Things in the world are natural individuals if, and only if, they are capable of experiencing phenomenal individuals.

The ontology implicit in the Central Thesis is a panexperientialist neutral monism. The fundamental kind is the causal nexus itself, and the nexus has multiple aspects: a phenomenal side, consisting of intrinsic properties that carry the components of the world's effective constraints, and an experiential side, to which the phenomenal natures are bound and through which they place their contributions to constraints. The Carrier Theory implies that neither experiencing nor phenomenal individuals are entirely physical because carriers are extrinsic within physics. They are nevertheless not epiphenomenal, nor do they interact with the physical. A variety of panexperientialism, as discussed earlier, also holds if the Central Thesis is true.

The discussions in chapters 5 and 6 show their bite here. Any natural individual is at least protoconscious: It is an experiential nexus even if it does not support thought. To avoid panexperientialism at this point, we would have to retract the proposed Central Thesis and assert a different form that covered the cognitive case (making it experiential) and that separately covered the rest of the world (making it nonexperiential) and ideally accompany it with an explanation of the discontinuity or continuity between the two. To characterize the two disjunctive conditions, we would have to overcome the obstacles discussed in chapter 6: We would have to explain why the experiential emerges in just the "right" contexts. The effort would invite a tremendously difficult theoretical problem whose result likely would be replacing the simple, straightforward Central Thesis given here with a much more complicated version.

The only motivation and the only payoff for that effort would be avoiding panexperientialism. How justified would the effort be, given just this motivation? I think earlier reflection has shown that this would be much effort for little return in the grand scheme of things. As the arguments in chapter 5 for the possibility of panexperientialism showed, we do not *know* that panexperientialism is false, so there would be no established facts driving the effort to complicate the Central Thesis.

At worst, panexperiential consequences are counterintuitive. Yet this is a fundamental theory, and science has already shown us in many ways—from the relativity, responsiveness, and surprising geometry of space and time to the randomness, state indeterminacy, nonlocality, and uncertainty principle of quantum mechanics—that commonsense intuition breaks down at the fundamental level of the world where the Central Thesis holds. Therefore, this kind of counterintuitiveness does not mean much when judging a fundamental theory such as the Carrier Theory of Causation.

Finally, it is not even clear in what sense the intuition against panexperientialism really is a *commonsense* intuition. Many other cultures have seriously entertained or endorsed an animistic metaphysics, and it is certainly possible that the current resistance to distantly related views such as panexperientialism is at least partly a knee-jerk reaction against these more primitive or theistic views. If so, the intuition against panexperientialism is not so much one of common sense but one of a specific cultural time and place. Perhaps some of it is rooted in a natural and admirable aspiration for sophistication as measured against a more primitive and superstitious past. However, we have to force ourselves to realize that there is nothing primitive or superstitious about the Central Thesis. Quite the contrary, in context it is a sophisticated proposal motivated by an unflinching adherence to modern standards of rational explanation.

It seems the reasons for outright rejection of the Central Thesis are weak. But why believe in the Central Thesis? Some strong reasons exist for adopting it. Whatever carries the nomic content of a natural individual must conform to the following:

1. They are intrinsic properties that are intrinsic *tout court*.
2. These properties must have the structural characteristics needed to carry effective and receptive properties.
3. The effective carriers must be determinables with the right kinds of internal contrasts among them, as well as relations of compatibility, incompatibility, and inclusion.
4. The receptive carriers must be neutral essences with a kind of inherent openness to their nature that can be filled by determinable properties.
5. Each of the receptive and effective carriers must have natures that are dependent on the nature of the other in the compositionally circular way that effective and receptive properties are dependent on one another.

In section 12.3, I preemptively discussed conditions 1 through 3 by defending the qualifications that make phenomenal properties good carrier candidates for effective properties. What of conditions 4 and 5? The experiencing subject is a good candidate for a receptive carrier that meets condition 4. In its normal state, the experiencing subject shows itself to be intrinsically plastic, suggesting a kind of neutrality, by binding and re-binding a vast variety of phenomenal properties, opening itself to a carnival of combinations and determinations of properties from the phenomenal world. Furthermore, the idea that experiencing is a kind of openness to phenomenal content coheres with common phenomenological reports about meditative states in which people are denied normal sensory input. In a physical state of sensory isolation, these meditative experiencers consistently report achieving a mental state that they identify as "pure" awareness in which consciousness is perceived as possessing a kind of contentless openness.

That leaves condition 5. Condition 5 is necessary because phenomenal properties, if they were just intrinsic *tout court*, lying next to one another in a Humean way, could not carry effective causation. The relationship between the effective

and receptive aspects of an individual must be metaphysically intimate. For properties to be *effective,* they must presuppose receptive connections as *positive* components in their own being and vice versa. In the relationship between effective and receptive causation, receptivity penetrates the being of effective properties, occurring as a presupposition in the very notion that they are effective. Furthermore, the logical intimacy between effective properties and receptivity plays an important metaphysical role. Through the intimacy of binding, the effective states of different individuals penetrate one another's being and present their constraints immediately. In a sense, having a shared receptivity provides a principle of substantial unity that activates the relations of requirement, compatibility, and incompatibility between effective properties, making these internal constraints between them relevant in specific ways to specific cases.

Plausibly, the ontological relation between phenomenal qualities and their participation in the experiencings of subjects matches this crucial logical structure of the relationship between effective properties and their shared receptivity. Focusing first on the phenomenal side, the phenomenal qualities of our consciousness seem to depend for their existence on entering into the experiences of a subject. Think of the paradox of unity. It is highly implausible, for example, that kinds of pain could exist for which there is no subject to experience them. If this is right, its possible role in experiencings is essential to pain. As for experiencing itself, claiming that something is an experiencing subject implies that it can experience phenomenal qualities. That is, its capacity to host and experience phenomenal being is essential to it.

Questions about the relations between the experiencing subject and its experiences raise many complicated and controversial issues. I do not have space to go into much here, but I do propose that phenomenal qualities could not exist unless some subject was experiencing them[7] and that experiences could not exist unless they were experiences of phenomenal qualities. Yet, despite this mutual participation in one another's natures, they are distinct essences. A phenomenal quality is an *object* of experience that should not be identified with the experiencing of it. And an individual experiencer is a subject of qualitative experience that should not be identified with its objects. So, just like effective and receptive properties, the experiencer and the experienced qualities constitute distinct yet interdependent aspects of the total individual.

Receptive fields and the content of experience. Recall that an individual's *receptive field* consists of the other individuals from whom it is directly receiving constraint. If the Central Thesis is correct, individuals experience the phenomenal carriers of the effective properties belonging to individuals in their receptive fields. Figure 12.3 can help us to visualize what this means for the experiencing of an individual.

> Definition 12.4: The *receptive field* of an individual $I_{n.k}$ consists of all the individuals $I_{n.x}$ . . . $I_{n.y}$ (1) with which it shares a receptive connection and (2) where it is on the receiving end of constraint with respect to that individual.

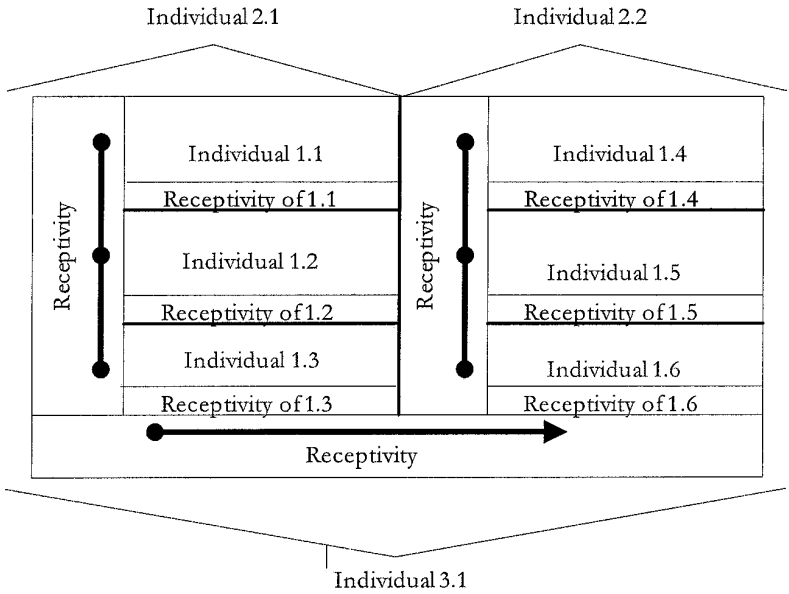


Figure 12.3 A representation of a complex individual. The highest level individual is at level three. It has an asymmetric connection between two level-two individuals, each of which symmetrically binds three level-one individuals. The receptive field of each individual includes the effective properties of the individuals directly constraining it. According to the carrier theory, an experiential property carries each individual's receptivity, and so the individual experiences the phenomenal properties that carry the effective properties in its receptive field.

Figure 12.3 depicts a compound level-three individual, labeled Individual 3.1. It consists of two asymmetrically connected level-two individuals, labeled Individuals 2.1 and 2.2. A beaded arrow representing the asymmetric receptive connection between these two individuals is drawn below them. The level-two individuals each consist of three level-one individuals. Two lines, each with three beads, are drawn within the receptive connections of these individuals to symbolize the symmetric receptive connections between their members.

By identifying these individual's receptive fields, we can work from the bottom level up and use the carrier theory to hypothesize experiencings for the individuals in the figure. At level zero, the individuals do not have receptivities of their own, so they could not have receptive fields and so could not experience.

Individuals 1.1 through 1.3 do have instances of receptivity belonging to them, and they also share a symmetric receptive connection within Individual 2.1. Because the connection is symmetric, each of the individuals in this nexus has the other two in its receptive field. For example, Individual 1.1 has individuals 1.2 and 1.3 in its receptive field. The effective state of each individual *realizes* one or more *effective properties* as described in chapter 10.

If the Carrier Theory of Causation proposed in this chapter is correct, these individual's effective properties are carried by phenomenal properties, and an experiential property carries each individual's receptivity. Individual 1.1's receptive carrier would enable it to experience the phenomenal carriers associated with the effective properties of individuals 1.2 and 1.3. Similarly, Individual 1.2's receptive carrier would allow it to experience the phenomenal carriers associated with the effective properties of individuals 1.1 and 1.3, and individual 1.3's receptive carrier would allow it to experience the phenomenal carriers associated with the effective properties of individuals 1.1 and 1.2. These relations are laid out in the table below.

Table 12.1

Experiencing Individual	*Individuals within its receptive field*	*Experienced qualities*
Individual 1.1	Individual 1.2, Individual 1.3	The carriers associated with the effective properties of individuals 1.2 and 1.3.
Individual 1.2	Individual 1.1, Individual 1.3	The carriers associated with the effective properties of individuals 1.1 and 1.3.
Individual 1.3	Individual 1.1, Individual 1.2	The carriers associated with the effective properties of individuals 1.1 and 1.2.

The analysis of Individuals 1.4 through 1.6 that are constituents of Individual 2.2 would be exactly similar. Notice, however, that the individuals within the two groups do not experience one another's carriers across the boundaries of their own nexii. Because carriers rely on transitivity to deliver their constraint, they only penetrate other individuals within the context of a shared receptive connection.

Level two presents an asymmetric case of experiencing. Individuals 2.1 and 2.2 are asymmetrically connected, and so Individual 2.2 is open to constraint from Individual 2.1 but not vice versa. Therefore, Individual 2.1 is in the receptive field of Individual 2.2 but not vice versa, and Individual 2.1 does not reciprocally experience Individual 2.2's phenomenal carriers. Finally, when we ascend to level three, we see that Individual 3.1 has the potential to experience, but, as the topmost individual, it is not bound within any higher level individual. To have a receptive field, and therefore to have phenomenal carriers to experience, it would have to be in a constrained slot of a shared receptive connection within a higher level individual. Therefore, it has no receptive field. Consequently the level three individual is not an experiencer.

We have nested experiencers here that look something like the Russian-dolls metaphor from chapter 4: There are experiencers within experiencers. However, the top- and bottom-level individuals do not experience. The bottom-level individuals (level zero) do not experience because they have no receptivity belonging to them, and the top-level individuals (level three) do not experience because

they do not belong to causal nexii from which they can obtain a receptive field. The individuals in the middle levels do experience.

12.5 Summary

Table 12.2 shows the requirements on the carrier role for natural individuals and how the experiencing of phenomenal individuals fills the bill. The next chapter discusses a host of more detailed and striking parallels between the observed character of experience and what one would predict for the character of carriers. For now, the high-level mapping goes like this.

Experiential property —> Receptivity

Phenomenal qualities —> Effective properties

Experiencing of qualities —> Reception of effective constraint

Experiential nexus —> Causal nexus

Table 12.2 partially details how the carrier relation is implemented.

Table 12.2 How the Experiencing of Phenomenal Individuals Fills the Carrier Role

Phenomenal properties (*feature of the proposed carrier*)	*Nomic content* (*structural requirement on the proposed carrier*)
1. The possibility of being experienced is essential to phenomenal qualities.	1. The possibility of being receptively bound is essential to effective properties.
2. Being an experiencing subject implies the experiencing of phenomenal qualities.	2. Being a receptive individual implies receiving the constraint of effective properties.
3. Phenomenal qualities are only potential unless actually being experienced.	3. Effective properties are only potential unless actually receptively bound.
4. Experience is only potential unless it is experiencing phenomenal quality.	4. A receptive connection is only potential unless it is binding effective properties.
5. Phenomenal properties are determinables.	5. Effective properties are determinables.
6. Phenomenological reports of the "pure" experiencing subject reveal a kind of contentless openness within pure consciousness.	6. Pure receptive connections are a kind of contentless openness.
7. Relations of inclusion, exclusion, compatibility, and incompatibility exist between phenomenal properties.	7. Relations of inclusion, exclusion, compatibility, and incompatibility exist between effective properties.
8. Scalar relations and relations of intrinsic difference exist between phenomenal properties.	8. Scalar relations and relations of stipulative difference exist between effective properties.
9. Despite mutually participating in one another's nature, phenomenal properties and the experiencing of them mark distinct essences.	9. Despite mutually participating in one another's nature, effective properties and the receptive binding of them mark distinct essences.

The Central Thesis solves the carrier puzzle quite neatly and fruitfully, and that is its best defense. It turns out that the causal nexus has three aspects: its effective dispositions, its receptive dispositions, and the carriers of this nomic content. The effective and receptive dispositional properties are the two complementary aspects of causation that give natural individuals their nomic content, and these two aspects are carried, ultimately, by the experiencing of interlocked subjects within the causal mesh.

Physical theory *specifies* some constraints between the effective states of the basic natural individuals by describing the nomic regularities that hold between their instantiations. One might argue (e.g., Stoljar 2001) that physical specifications indirectly *designate* phenomenal properties that are involved in carrying these effective constraints. When combined with the Theory of Causal Significance and the Carrier Theory, physical theory *suggests* the existence of protoconscious experiencers at many levels of nature.

The Central Thesis does have a price of admission, and that price is its implication that some kind of panexperientialism is true of our world. Just how widely spread experience is remains to be discovered, as the question of which individuals are the natural individuals is a substantial and important scientific question. However, as I argued earlier, we should have expected to arrive at some kind of panexperientialism, and this kind of panexperientialism is a *benign panpsychism* because experience is likely to be very simple in the vast majority of cases, to be restrained to highly specialized circumstances despite its outrunning cognition and to be qualitative content unaccompanied by thought whenever it exists outside of cognitive contexts. Putting panexperientialist implications aside, the final question is just this: How does all this help with understanding human consciousness?

13

The Consciousness Hypothesis

13.1 Consciousness and High-Level Individuals

On the sea of individuals, human beings are tidal waves. At least that is the most natural conclusion if one accepts the Central Thesis as the best solution to the carrier problem. Recall the Central Thesis from the last chapter:

> *The Central Thesis*: Things in the world are natural individuals if, and only if, they are capable of experiencing phenomenal individuals.

The Central Thesis is an informal way of stating an axiom schema whose instances describe fundamental facts at all the levels of nature. It naturally leads to the Consciousness Hypothesis, which is:

> *The Consciousness Hypothesis*: Each individual consciousness carries the nomic content of a cognitively structured, high-level natural individual. Conscious experience is experience of the total constraint structure active in the receptive field of such an individual.

The Consciousness Hypothesis brings this book full circle. After being led to causation by the problem of consciousness, we rejected conventionalist/Humean views, developed a substantive alternative, and raised the carrier problem. The Central Thesis solves the carrier problem by proposing that phenomenal experiencings are of carrier content in the receptive field of a natural individual, thereby taking the discussion from causation back to experience. However, the kind of experiencing required could just be the kind of simple, precognitive experiencing entertained by panexperientialism: pure feeling too simple to support anything worthy of the name "consciousness." The Consciousness Hypothesis is an application of the Central Thesis that lifts experiencing up to full-fledged consciousness.[1] With it, the circle is closed.

In support of the Consciousness Hypothesis, this chapter reviews the issues raised in part I. It discusses directly how the proposal here allows Liberal Naturalism to avoid the explanatory failings of pure physicalism. It also shows how it provides explanatory success by solving the puzzles, paradoxes, and tensions confronting Liberal Naturalism. Just as the preeminent virtue guiding construction of the Theory of Causal Significance was simplicity, and as the Carrier Theory of Causation raised intelligibility and uniformity into position as equal partners, the Consciousness Hypothesis is an attempt to show fruitfulness. I end the chapter by discussing how this fruitfulness supplements the philosophical reasons earlier introduced for accepting the existence of receptive connections and provides additional support to the model.

13.2 Avoiding the Failures of Physicalism

Chalmers (1996) has nicely summarized the five extant arguments against the logical (or ontological) supervenience of consciousness on the physical. They are (1) the logical possibility of inverted spectra, (2) the logical possibility of zombies, (3) the epistemic asymmetry between facts about consciousness and other facts, (4) the knowledge argument, and (5) the absence of analysis. I consider these arguments, as they affect Liberal Naturalism, in reverse order.

The absence of analysis. The key premise of the argument from the absence of analysis is that, for a property to ontologically supervene on the physical, it must be at least roughly analyzable into categories that the physical facts might entail. The antiphysicalist argues that there can be no adequate analysis of phenomenal properties into the relevant functional and structural terms, so it seems that consciousness must not ontologically supervene on the physical.

The Liberal Naturalist's Consciousness Hypothesis is not a reductive hypothesis, so that kind of conclusion has minimal force against it. According to the Carrier Theory of Causation, the primitive carriers of nomic content must meet certain general conditions: They must be properties that are intrinsic *tout court* (i.e., not intrinsic to any system, in the technical sense discussed in the last chapter); they must be extrinsic within physics; they must have internal contrasts that mirror the stipulative contrasts they carry; and they must be characterized by a structured interdependence that appropriately mirrors the compositional circularity between effective and receptive properties.

An analysis of experience suggests it meets all these general conditions. Still, liberal naturalists do not claim that The Central Thesis represents an analysis of consciousness without remainder and so they do not claim it reduces consciousness to something else. Instead, it accepts without qualification the existence of intrinsic information available only through acquaintance. Liberal Naturalism is in fact welcoming of the non-analyzable aspects of conscious experience because the liberal naturalist justifies the Consciousness Hypothesis partly on the grounds that it finds a useful nonreductive place in nature for the otherwise extraneous information that acquaintance acquires.

The knowledge argument. Even in principle, a person with perfect physical knowledge could not use that basis of physical information to derive the phenomenal information available in experience. Proponents of the knowledge argument claim that this shows that the phenomenal facts are extra facts, over and above the physical. The Carrier Theory of Causation implies that full knowledge of an ultimate carrier requires acquaintance. By the Carrier Theory's own lights, Mary could not have full knowledge of the carriers from inside a black-and-white room.

Proof: Imagine that Mary is a brilliant neuroscientist locked in a black-and-white room, and that she wishes to know everything there is to know about causation in our world. Her knowledge would have to include knowledge of the receptive structure of our world and of the carriers. She could conceivably theorize about the receptive structure given the effective facts of natural science, great talent with inference to the best explanation, and a little luck. From this, she could infer the structure of internal contrasts that hold between the carriers. However, these are all systematic facts, so she would of necessity still be missing some facts about the carriers, the facts corresponding to their intrinsic character. In the terminology of the last chapter, these facts about intrinsic character are facts about things that are *extrinsic within* the system that Mary otherwise has perfect knowledge about. All theoretical knowledge is discursive and systematic, so the only way for her to get these facts would be acquaintance with the relevant intrinsic natures that fill the carrier role in her world. Hence, the Consciousness Hypothesis implies that for Mary to have *all* the facts, she would have to have some further experiential facts about intrinsic natures. The knowledge argument against the Consciousness Hypothesis fails.

Epistemic asymmetry. Our reasons for believing in the physical facts and in other facts ontologically supervenient on them are straightforwardly based on external evidence. If consciousness ontologically supervened on the physical, external evidence would give us reason to believe in it. However, the external evidence does not give us adequate reason to believe in consciousness. Our only reason for believing in consciousness is the first-person fact that we ourselves are conscious, our direct experiential acquaintance with it. The argument from epistemic asymmetry suggests that a physicalist's explanation of consciousness inevitably would fail to account for some of our evidence about consciousness. Worse, the evidence for which it would fail to account is precisely the evidence responsible for our belief in consciousness in the first place, which is entirely unacceptable.

The Consciousness Hypothesis is an instance of the Central Thesis, which is an axiom schema. The Liberal Naturalist's justification for the Central Thesis partly relies on internal evidence, qualifying the experiencings of phenomenal individuals for their role as ultimate carriers based on information available only from first-person experience. The analysis in no way suggests that the carriers are reducible to something else that we believe in purely based on external evidence.

Because the force of the argument from asymmetry is just that any facts failing to evade it will have to be included based on internal evidence, the Consciousness Hypothesis meets its demands.

The logical possibility of zombies. A zombie would be a being physically identical to you or me yet lacking subjective experience. The logical possibility that our functional structures might not produce consciousness if realized in nonorganic materials implies that even facts about our own organic physical structures could be true consistent with the absence of experience. There is, after all, no more of a conceptual connection from organic chemistry to consciousness than there is from other physical structures to consciousness. This strongly suggests that a zombie world is consistently conceivable and therefore possible (given an appropriate analysis of the link between conceivability and possibility).

There is no exactly analogous argument against the Consciousness Hypothesis because experiencings are built into the fundamental nature of the world. To be even roughly analogous, a zombie argument against the Carrier Theory might try to establish the logical possibility of a world in which all the facts about its effective and receptive properties are the same as in our world but in which there are different carriers and so no consciousness. Such a world would at least have the same nomic content as ours, and so its possibility would show there still remains a certain kind of contingency surrounding the facts of experience.

However, the discussion of the knowledge argument makes it plain that no one can positively conceive of alternative carriers in the way needed to justify the logical possibility of such a world. In particular, a conceivable world is logically possible just in case its conception is consistent when the intensions on the concepts are made *suitably definite*. To be suitably definite, the intension must enable recognition of the reference for the concept when evaluated within a possible world.

In the purported zombie world, the intension would have to uniquely pick out a set of intrinsic properties which are not experiential properties but which nevertheless carry cognition. But consider how we conceive of intrinsic properties with which we are not acquainted, such as Mary's conception of phenomenal redness before experiencing it. Her concept was indirect, having content only deferentially, so she was in position to conceive of phenomenal redness only as an intrinsic property like "the one with which other people are acquainted." But deferential concepts will not provide a conception of alternative carriers in a zombie world because no one exists to which we can defer. To conceive of a world with nonexperiential carriers, someone would have to have acquaintance with the appropriate kind of intrinsic nature, and no one is in that position.

In the zombie case being proposed, first-person ostension cannot be the basis from which to bootstrap a conception, and it is very unclear how else one might get the requisite concepts. Therefore, it seems that we cannot successfully conceive of worlds with alternative kinds of carriers and so cannot conceive of a Liberal Naturalist zombie world. Certainly, the burden is on someone who claims to be able to conceive of carriers that are not phenomenal individuals. They at

least must be able to convey to *other* people what they have in mind. At best, there is a kind of prima facie negative conceivability in which we cannot rule out nonexperiential carriers for a high-level, cognitively structured individual, but prima facie negative conceivability cannot deliver strong conclusions about the possibility of a world (Chalmers 2002).

Even so, arguing that we cannot positively conceive of a zombie world is not the same thing as arguing successfully against the possibility of such a world. All we know is that the zombie world is not humanly positively conceivable. To show that the world was not possible, we would have to show that it is *in principle* contradictory. We cannot do that either. The question remains, Could there have been carriers that are not experiencings of phenomenal individuals? That seems to be an open question.

In my conservative moments, I want to deny it. Invoking Occam's razor, I remind myself that we cannot conceive of anything meeting the description of ultimate carriers except phenomenal individuals, and we should therefore conclude that these are the only things that *could* do the job. This is the simplest answer because it avoids raising any further questions.

In my more sober moments, I believe that we cannot rule out the possibility of alternative carriers and should, charitably, allow them. I treat the phenomenal facts about our world as contingent and implementational in character. If that is the case, they could be substituted for without changing either the effective or receptive structure of the world. So in a sense zombies are probably possible, even if we cannot conceive them, but their possibility does not affect the truth of the Central Thesis or the Consciousness Hypothesis.

In my humble moments, I am just agnostic. I recognize that this is a question about the world-making ingredients that God might have had available in the jars of his kitchen cabinet. In the face of such awesome questions, we should simply turn away.

The inverted spectrum. We can consistently conceive of a world in which the physical facts are the same but in which color perception is systematically inverted. For instance, if a creature sees in grayscale, we can consistently conceive the black-and-white axes being switched.

In their original context, inverted spectrum cases work only against reductive accounts of consciousness. Under the Consciousness Hypothesis, phenomenal colors constitute base facts rather than reduced facts. The reduced facts are the facts about the structure of effective and receptive causation, and the reduction is to facts about experiential subjects. Standard inverted spectra cases, even if they went through against the Consciousness Hypothesis, would just show that states of effective causation may be multiply realized. By itself, that conclusion is metaphysically harmless to the Consciousness Hypothesis and Liberal Naturalism. It would present some epistemological danger, but that danger could be avoided by appealing to simplicity constraints that support a suitable hypothesis about the uniformity of nature.

13.3 Resolving the Puzzles, Paradoxes, and Tensions

Part I of this book explored a series of puzzles, paradoxes, and tensions brought on by the rejection of physicalism. One tension was the threat that panexperientialism is a likely outcome of the Liberal Naturalist turn. Beyond that threat were six further puzzles and paradoxes: a puzzle about the unity of consciousness, a paradox about the simultaneity of the subjective instant, a puzzle about the epistemology of consciousness, a puzzle about its seeming superfluity, the paradox of the grain problem, and the tension of the boundary problem for phenomenal individuals. Now I return to these problems, suggesting ways that the Consciousness Hypothesis may help to illuminate or resolve each of them. The end result is the striking discovery that the fundamental carriers of effective and receptive causation would have predictable properties that parallel the troublesome properties of consciousness.

By shedding light on mysteries surrounding consciousness, this chapter section will also bolster the case for the existence of receptivity and the plausibility of the Consciousness Hypothesis. After all, although the previous chapters described the *role* of receptivity some readers may have a gnawing sense that they still do not have a good idea of what a receptive field really is and why they should believe in it. The basic discomfort may be that, despite the formal explanation, receptivity, receptive fields and carriers still seem somehow odd or alien to our experience of the world. The following discussions collectively provide evidence that this feeling might be mistaken.

Ubiquity and fundamentalness. Earlier arguments set an expectation that whatever the qualitative field turned out to be, understanding its basis would (1) help us to "get under" physics by showing a way to see the physical and experiential as coequal aspects of a deeper kind, (2) show it to be surprisingly widespread, and (3) show it to be fundamental. The pillars of the Theory of Natural Individuals collectively meet all three of these expectations.

The Carrier Theory "gets under" physics because it provides a categorical causal basis and intrinsic character to a world that is incompletely and schematically described by physics. The Central Thesis implies that experience is surprisingly widespread: The qualitative field is in reality the receptive field of a natural individual. Natural individuals occur at several levels of nature, occurring at least from the microphysical to the human level, although the totality of things that are natural individuals is still unknown. Finally, specific carriers such as consciousness are fundamental, being introduced within instances of the Central Thesis, an axiom schema for the theory.

The unity of consciousness. The problem of the unity of consciousness has roots in an elusive intuition. The experiential elements of consciousness do not seem, intuitively, to be capable of independent existence in the same way as proper components of a system are. Rather, their existence as elements within the experiential manifold seems to presuppose the existence of the experiential manifold

within which they are elements. I suggested that the unity of consciousness is somehow produced by the holistic dependence of each element in this way on the existence of a common whole.

The problem of the unity of consciousness presents the Liberal Naturalist with one challenge and one puzzle. The challenge is to articulate the unity of consciousness more clearly and precisely. The puzzle is to explain how a single system, as consciousness, may have this kind of unity as observed introspectively and, as brain, not have this kind of unity as seen from the outside.

We can understand the unity of consciousness on the model of the compositional circularity holding between the receptive and effective properties. Recall from the earlier discussion of diagram (g), figure 10.13, that individuals may have effective *states* outside of a causal nexus, but that these states only present effective *constraints* within the context of a nexus. That same discussion explained why an individual's effective properties should be identified with the constraints it presents rather than with its effective states. From these two facts we can deduce that only in the context of a nexus do natural individuals realize effective *properties*. Therefore, outside of the receptive context of a causal nexus, no effective properties exist. From this and a simplicity assumption, we can further deduce that no *carriers* of effective properties would exist outside of a receptive context. Phenomenal properties, because they carry effective properties, should be brought whole into existence and leave existence with the individual fields of effective constraint they help to constitute.

This dependence of effective properties on binding within the causal nexus is the by-product of the compositional circularity in the natures of effective properties and receptivity, an interdependence that means that each element needs a causal nexus to exist before it can gain a completed and determinate nature. This deduction resolves the challenge presented by the unity of consciousness: If the Consciousness Hypothesis is true, we can clearly articulate what it means to have the kind of unity that consciousness exhibits by appealing to the metaphysical dependence of effective properties on causal nexii for their completion and realization.

Having addressed the challenge, we are in position to address the puzzle: A conscious system may have the kind of unity that consciousness seems to have introspectively because an actually existing effective property is identified with the contribution it makes to the constraint structure imposed within a causal nexus. These constraint structures come to exist only within receptive experiencings of their carriers.

The other part of the puzzle is to explain why that same system would seem *not* to have that kind of unity when viewed as a brain by outside observers. This second part of the puzzle is solved by recalling the *partially* reductive character of the hierarchy of individuals in the causal mesh (figure 9.12). No individual above level one is wholly reducible to the lower level individuals bound within it, as each receptivity is unique to the individual it helps constitute. Nevertheless the bound individuals within the nexus do not depend on the existence of the

nexus. From the perspective of a third-party observer, the high-level receptivity of the individual would be far less striking than the hierarchy of individuals it directly and indirectly binds.

Together, these observations promise to solve the puzzle. The nonreductive element of the individual, its receptivity, is responsible for the system possessing the appropriate unity in its phenomenal character. A receptive connection facilitates the realization of effective properties (and thus effective carriers) by providing an appropriate context in which those effective contributions can be carried. The system's hierarchy of lower-level individuals, which is its reductive aspect, appears to be a component system to outside observers. Finally, notice that this is an explanation of phenomenal unity only and that the existence of this kind of phenomenal unity is compatible with various kinds of functional disunity of consciousness that have been observed, such as split-brain disorders, schizophrenia, and dissociative personality disorders.

The subjective instant. The conscious subject occupies a kind of privileged reference frame in which conscious events are all occurring simultaneously. Yet we know that conscious events correspond to asynchronously occurring brain events, events for which there is no privileged reference frame. The problem of the subjective instant requires reconciling these two facts.

Fully solving this problem would require a separate treatment of space and time such as the one sketched in chapter 10, section 6. Here I can only show the outlines of what a Liberal Naturalist solution could look like given the earlier described approach to the larger problem. The key lemma in that proposal was that spacetime is not primitive, being reducible instead to more fundamental facts about the world's receptive structure. In the proposed model, natural individuals provide frames of reference for constructing spacetime, and distance in space and time between events is a projection of dependency and immediacy of interaction between individuals.

From the proposed model we know that spacetime projects from a basis that is layered vertically to reflect the hierarchical nature of the causal mesh, as well as being organized horizontally in each layer. Within this vertical scheme, the state of a higher level individual must have a dual character. It would have to be, at once, an immediate *single* state determination for the higher level individual and a *multiplicity* of state determinations for the lower level individuals bound within it. That is, a higher level individual's irreducible state S would be a function, $f(x,\ldots,z)$, of the states of a multiplicity of bound individuals. The mystery of the subjective instant can be clarified by understanding why, when the higher level individual (with its experiencing receptivity) is chosen as the frame of reference for determining a spacetime mapping of events, projecting this situation into a coherent spacetime scheme could require mapping the instantaneous state for the higher level individual onto a duration of states in the existence of the lower level individuals.

Consider a causal process like that depicted in figure 13.1. The process con-

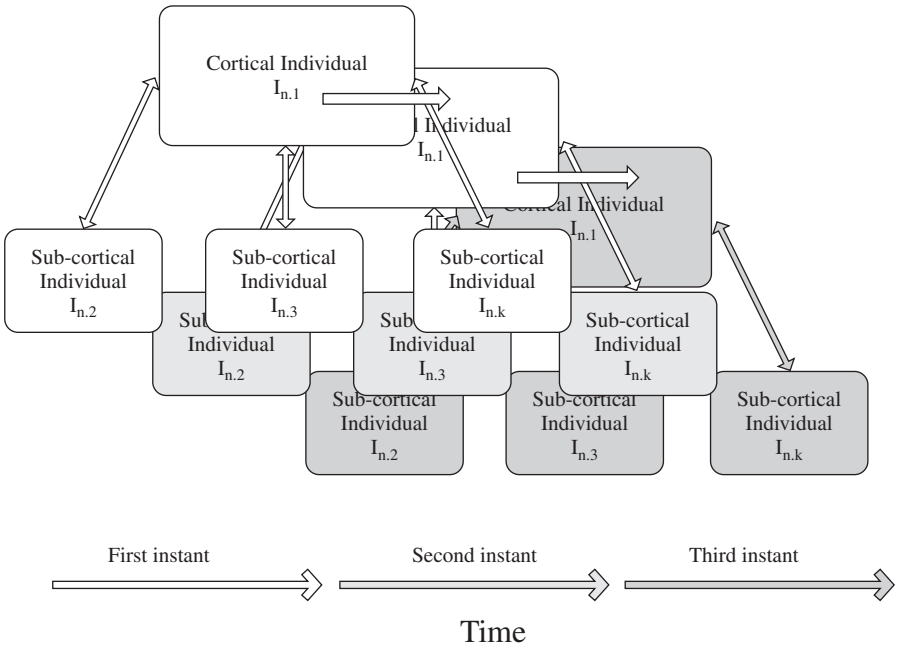


Figure 13.1 One possible model of the receptive structure of human consciousness. The model depicts a process of high-level individuals. Here, the stream of human consciousness would consist in the experiencings of the successive higher level individuals in the cortical process. The two questions relevant to understanding the paradox of the subjective instant are these: How would this individual experience the temporal character of its subjective state? How would it construct an objective representation of time?

nects a series of higher level individuals $I_{N.1}$, $I_{N.2}$, and $I_{N.3}$ into a process and possesses the following structure. It consists of two asymmetric connections depicted as one-way arrows connecting $I_{N.1}$ to $I_{N.2}$ and $I_{N.2}$ to $I_{N.3}$. Additionally, assume that each individual $I_{N.1}$, $I_{N.2}$, and $I_{N.3}$ is itself part of a symmetric nexus involving several sub-cortical individuals and in which each member is open to the constraining presence of every other member. Finally, the overlaps between them represent that $I_{N.1}$ shares some of its members with $I_{N.2}$, which shares some its members with $I_{N.3.}$

I put figure 13.1 forward tentatively but not arbitrarily. Conscious states are plausibly members of causal processes (a *cascade* in the spacetime terminology of chapter 10, section 6), and the kind of causal process depicted in figure 13.1 provides a plausible template for understanding the high-level intrinsic causal structure of a conscious mind. In this template, each receiving member of an asymmetric connection, such as $I_{N.2}$ in the connection $I_{N.1} \Rightarrow I_{N.2}$, would experience the carriers of a highly complex set of constraints placed by a constraining

member such as $I_{N.1}$. The constraining member would achieve its complexity of constraint by itself being a highly complex natural individual.

As explained in figure 10.13, diagram (g), an individual with such a complex effective state could produce an enormous number of subtly characterized and simultaneously occurring effective properties. Additionally, because overlapping individuals in the higher level process may partially share members (as elaborated throughout chapter 10), the constraints associated with these members may come to exist at multiple moments in the subjective time of the process. From this we can deduce that any such elements within experience would have *duration*.

From the reference frame provided by each individual $I_{N.\mathrm{k}}$, its own state determination would constitute a subjective instant: It is the result of receiving a single unified field of effective constraint and its reception is a single action on that field. Any elements of experience which co-exist within its field of constraint would exist simultaneously within its experiencing of that field, and they would continue to exist simultaneously within the experiencing of the processional cascade for as long as their durations overlap.

Furthermore, these are objective facts about time for the subject: The individual's receptivity would provide a kind of privileged reference frame on these elements of its state determination because it represents the subject's point of view on a set of high-level immediate interactions. We can deduce that, by relativizing time to this vertical dimension of reality, one relativizes the concept of an *instant* and of what counts as *instantaneous* along a dimension not recognized by relativity. A *vertical instant* is relative to immediacy of causal interaction between individuals at a level of nature. It can be identified objectively with the reception of a unified structure of effective constraint. For the experiencing subject, its irreducible receptive connection does provide a privileged frame of reference from which to experience this immediacy of interaction.

However, the causal situation contains more than the immediacy of presence that is revealed to it experientially. Knowledge of the total situation reveals an internal structure to it and to the members of its receptive field that is normally hidden to the higher level individual. When fully articulated, its members' internal structure could be very complex, involving many, many layers, with corresponding relations of orderliness and fixity. For example, each of the member individual's bound by the individual $I_{N.k}$'s could be asymmetric nexii possessing yet further internal structure themselves. These relations of nesting and overlap can repeat for many, many levels down to the microphysical basis of nature. If the Consciousness Hypothesis is true, then some higher level natural individuals are cognitive. How could these complicated, layered individuals be organized into a spacetime projection from an arbitrary third-person perspective?

There is a plausible case that a paradox would result if some of these cognitive individuals tried to internalize a third person construction of time, viewing themselves "from the outside" without realizing the relative character of vertically instantaneous states. We know that as their observations moved *downward* in the hierarchy of natural individuals, their understanding of the states of individuals

like themselves would show them as more and more internally complex. These highly complex individuals would be essentially impossible to compress into instants from *other* frames of reference within an intellectual model where time is itself flattened into one metric applicable at all levels. Under that kind of cognitive demand, what seems to be an *internally* instantaneous state of a high-level individual, as experienced by the individual itself, might be fruitfully reconstructed only as an *external* duration when projected into spacetime from an arbitrary frame of reference using more complete information about all of its layers of internal structure.

It seems that as cognitively structured individuals gained knowledge of their many layers and as they try to assimilate the evolution of all layers into a single flat-structured temporal framework, the internal complexity of the higher level component individuals would force a projection of their states onto durations at the lower levels. So, as the temporal hierarchy of the mesh is flattened, the individuals involved in higher level processes should become squeezed and forced to take on temporal depth that is not necessarily apparent from the inside. They would therefore find themselves losing the instantaneous character of their own states, internally observed.

The knowledge paradox. The knowledge paradox stems from our knowing that we are conscious even though experiencing seems not to be causally *responsible* for our brain states. On the carrier view, one might have a nagging feeling that causal responsibility hangs only off the effective side of things. One might worry that the phenomenal qualities can be omitted from adequate explanations of behavior, and so their job of carrying the effective constraints may not earn them causal responsibility.

The question of what kinds of things may claim causal responsibility in our world is subtle, so we should avoid making too much of these kinds of worries. Imagine that Trey has a date with Carol and that she is supposed to meet him for dinner at Everybody's Pizza on Wednesday. At dinnertime on Wednesday, Trey is waiting for Carol, but Carol fails to show. Trey sits at his table, waiting, and as time goes by he realizes what has happened. At first he feels disappointed, and as he reflects on it he gets angry. Then he leaves.

In this situation, what caused Trey's feelings of disappointment and anger? I suggest that, at least at first blush, his feelings were caused by *Carol's not showing up*. Notice what a strange sort of thing that is. It is a negative. It is nothing with which Trey's brain states interacted (there was no exchange of energy between Trey's perceptual apparatus and Carol's failure to show). It was not even an event, unless there are such implausible things as negative events that make negative facts true. The thing causally responsible for those emotions, it seems, was an absence. This suggests that causal responsibility accrues to things without physical causal powers.

The puzzle case of Trey and Carol's broken date throws some doubt on the thesis that only things with physical causal powers can have causal responsibility, but

I will not use this example to argue for a final conclusion on the causal responsibility of negative facts. Instead, I consider the two most likely ways of responding to the puzzle case, and I argue that either position one takes on it leaves an opening through which we might possibly ground knowledge of consciousness.

Case 1. The negative fact that Carol did not show up caused Trey's disappointment and anger. If so, the cause of Trey's anger is the negation of a fact. But the negation of a fact is not a physical event, so this implies that causes need not be physical events (and, by implication, physical properties or facts). A cause of an event may be a structural fact about the world involving *abstract objects,* such as the closure conditions necessary to draw negative conclusions.

The world's receptive structure, the Liberal Naturalist submits, is also a structural fact, although of a different sort, and no less able to bear the weight of causal responsibility. The carrier facts are about the implementation of that structure. These two kinds of facts are therefore constitutive and structuring causes of our behavior. This relationship, the Liberal Naturalist further submits, is intimate enough to be justificatory. At the very least, we need a solid argument that these facts, as causes, are problematic in a special way that negative facts are not. Only then do we have reason to worry that facts about consciousness cannot be justificatory.

Case 2. The negative fact that Carol did not show up has no causal responsibility. In its place, we may substitute something like the presence of the unfulfilled expectation. The unfulfilled expectation may be a physical state of the central nervous system and, therefore, a positive fact with physical causal power. However, the negative fact still shows up as the *reason* that the expectation went unfulfilled. This answer commits us to the position that *reasons do not need to be causes*.

The negative fact, as a reason for Trey's feelings, still plays a *justificatory* role in establishing them as proper in his situation. At the first step, it still helps to *explain* his feelings, and, at the second step, it is crucial to establishing those feelings as *proper* in the overall situation. It follows immediately that the possession of certain mental states may be *justified* by facts that need not be causes of those mental states, either directly or remotely.

With consciousness, the facts about receptivity and carriers will certainly count as reasons for our mental states. After all, even if they do not earn causal responsibility, they are crucial components supporting our total causal situation (e.g., had these carriers and receptive facts not existed, these mental states would not have existed, either in their physical or phenomenal aspects). By having a place in the full explanation of the existence of our beliefs, occurring as constitutive and structural reasons for our having the brain states that we do, the presence of that structure and those carriers should be able to play a justificatory role in the full story of why we have the mental states that we do.

Whether we allow negative facts to have causal responsibility or deny that they can have it, both cases leave some wiggle room for us to have knowledge of

consciousness. Even so, the Liberal Naturalist still has the problem of giving a positive account of the epistemology of consciousness, and that, like epistemology generally, moves fast into murky area. Although we can see that there is no special problem of the causal irrelevance of consciousness, one may still wonder just how the positive story goes.

I can see one helpful element of any potential solution. First, I believe a proper solution requires recognizing that we have a third type of knowledge over and above the propositional knowledge expressed by "knowing that . . ." clauses and the skillful knowledge expressed by "knowing how . . ." clauses. This third kind of knowledge is a kind of empathic knowledge expressed by "knowing what . . ." clauses. Examples of this kind of knowledge are: knowing what it is like to hear a scream, knowing what it is like to smell a baby, knowing what it is like to have something on "the tip of your tongue," knowing what it is like to fear death, and so forth.

This "knowing what . . ." is a kind of knowledge by acquaintance, and it is not truth evaluable, just as skills are not truth evaluable. Instead, it is a basic way of being for conscious subjects and is presupposed by the other kinds of knowledge. The epistemic puzzle for consciousness does *not* concern how we may have knowledge in the "knowing what . . ." sense. This knowledge is knowledge of the basic causal nature of the particular that we are. It is acquaintance with the carriers of our own nomic content and is available to us because of the immediate nature of the shared receptive connection that consciousness carries.

The problem is to explain how this "knowing what . . ." justifies instances of propositional knowledge expressed by "knowing that . . ." clauses. How does the intimacy of acquaintance license the uttering of sentences? This is a deep puzzle for epistemology generally, I believe, and does not arise specifically for consciousness alone. I would guess that its solution lies in giving some epistemic value to the particularity of a creature's circumstances and trying to understand the *variety* of ways that particularity is responsible for that creature being what it is, in the states that it is in. Beyond this, I will propose that a proper theory of representation involving certain kinds of *knowing how* in an essential way can act as the needed bridge between *knowing what* and *knowing that*.

Once these three kinds of knowledge are all recognized and explained, the knowledge paradox for consciousness reduces to the problem of explaining how instances of *knowing what* can support instances of *knowing that*. As I mentioned in chapter 5, I favor action-oriented views of representation in which the representational content of a type or token is determined by the way it provides guidance for a subject's action. I give the details of the theory elsewhere (Rosenberg and Anderson, 2004), and if we assume that something like this guidance theory is true, it is possible to speculate in an interesting and substantive way regarding how *knowing that* can emerge from *knowing what* with the help of *knowing how*.

Specifically, to solve the knowledge paradox we need an explanation of how a subject can have a representation with the peculiar properties required by our

representations of conscious experience. Note that Chalmers (1996) distinguishes between three kinds of knowledge that we express about conscious experiences:

1. *First-order phenomenal judgments.* Judgments that occur via conscious sensations but that concern the object of an experience rather than the conscious sensation itself. For example, "that shirt is purple" And "this soup is hot" are first-order judgments.
2. *Second-order phenomenal judgments.* Judgments about the occurrence of sensations and qualities of experience themselves. For example, "I am feeling a very sharp pain" and "I see a particularly strong shade of red" are second-order judgments.
3. *Third-order phenomenal judgments.* Judgments about conscious experience as a type. For example, "consciousness exists," "phenomenal qualities are not structures of bare difference," and "phenomenal red is a warm color" are third-order judgments.

Phenomenal judgments express representational content. The properties of these representations for which we need to account are:

1. These representations can in fact support first-, second-, and third-order judgments about conscious experience.
2. With respect to first-order judgments of consciousness, they are ordinary fallible judgments.
3. The representations expressed by second- and third-order judgments are about our conscious experience and elements of conscious experience.
4. With respect to second-order judgments of consciousness, the representations allow that some second-order judgments about conscious experience are instances of certain a posteriori knowledge, while still allowing that we are in general fallible about second-order judgments of consciousness.
5. With respect to third-order judgments of consciousness, the representations support many instances of certain a posteriori knowledge and far fewer instances of fallible knowledge.

To rein in the problem, we can start at the bottom level by coming to understand our *knowing what*, which is knowledge by acquaintance. From there we may be able to work our way up from first- to third-order phenomenal judgments.

Knowledge by acquaintance. This chapter and the previous chapters on the foundations of causation provide ways for us to model our acquaintance with phenomenal properties. The Consciousness Hypothesis tells us that conscious experience results from the existence of a natural individual capable of cognitive processing. From The Central Thesis, we can deduce that this individual is receptively experiencing a structure of constraint carried by a unified manifold of phenomenal properties. Current evidence (see chapter 14) points to a likely scenario in which these carriers carry the vector codings fed into cortical systems by more primitive systems, along with codings of the previous states of the cortical

system itself. The thalamus seems to play the most critical role as a mediating system, with central help from the hippocampus and the limbic system.

This current research suggests that the natural individual we identify as our conscious selves is a cortical individual. We can deduce from its reception of carriers that it and those from whom it receives effective constraint share a binding within a single causal nexus. *Binding* here refers to the metaphysical relation introduced in chapter 9, section 11 (not to the neural process with the same name), and it enables aspects of different individuals to enter into one another's nature. Acquaintance should be identified with phenomenal properties becoming part of the intrinsic nature of a cognitive (likely, cortical) natural individual that is receiving them. There is thus no appreciable metaphysical distance between the experience of the phenomenal property and the thing experiencing it. If the Consciousness Hypothesis is true, binding as developed in the Theory of Causal Significance metaphysically underwrites *acquaintance*.

Because acquaintance collapses the metaphysical distance between thing experienced and thing experiencing, it raises the possibility of a mechanism that can bootstrap this metaphysical situation into a collapse of the epistemic distance between phenomenal thing known and propositional knower. Clearly, this does not happen with first-order phenomenal judgments. First-order judgments of the sort, "that shirt is purple," are not really about conscious qualities at all. They express representations whose content is about (purported) public properties of public objects. These judgments are as fallible and nonmysterious as any other sort of representation with public content. What makes first-order judgments of the type, "that shirt is purple," interesting is not their representational content but the representational vehicle: *phenomenal purpleness*. By virtue of carrying the effectiveness of vector codings that become bound to the cognitive processing of a cognitively structured individual, phenomenal purpleness itself becomes caught up in the representation-consuming activity of the cognitive engine. Within a cognitive context, therefore, phenomenal properties advance from being simple carriers of effective constraint to carriers of representational guidance. In gaining this higher level property of *having representational content*, they create the potential for useful further adaptations that take advantage of this content.

In their capacity as representational vehicles for first-order judgments, phenomenal properties provide guidance to the organism by tracking purported features of the environment. To perform this function, the phenomenal properties themselves must have interesting features whose presence can be detected by the subject's action-oriented processing, whose variances can be read by that processing, and whose structure and variances reliably track environmental features. From a design standpoint, there are circumstances in which a subject with the ability to track its own representational vehicles would have an advantage over a subject that did not have the ability.

One circumstance in which an ability to track its own representational vehicles would be useful would be one in which the representational vehicles became corrupted in some way. If they became corrupted, their features might be less useful

for tracking environmental features than they normally would be, and knowing this could be helpful to the subject. Blurry vision might be an example of this kind of tracking. Another circumstance in which it could be useful would be one in which the vehicle became disconnected somehow from other representations, so that it was no longer possible for the subject to usefully correlate its tracking information with other tracking information provided by other representations. An example of this would be a sound that is heard but not perceptually placed as coming from any specific direction or object.

Given that representational vehicles are already directly bound into the subject, the elegant and most reliable way to track them would *not* be to create second-order representational vehicles that track the first-order representation vehicles. This would be wasteful of processing and unreliable, as it would create the opportunity for second-order error. The elegant solution is to design a second processing mechanism that uses the very same representational vehicles, the phenomenal properties, as self-representations. This kind of trick, using two different interpretational mechanisms to extract multiple meanings from one semantic vehicle, is used by natural selection elsewhere, including at the ground level of biology in the decoding of DNA. It is also a common trick within computer science, where it is not unusual to have different procedures that treat the very same data structure in two semantically distinct ways.

There are several engineering advantages to using an entity as a representation of itself. One advantage is that the solution uses fewer resources. Instead of having to develop a new decoder and a set of new, second-order representational vehicles, a system needs only a new decoder. Another advantage is that it eliminates a source of potential error, as the possibility that the representational vehicle will be off track with respect to its content is eliminated. This isolates the possibility for error in the decoding mechanism. A third advantage is that it bypasses the logical regress that can arise from the desire for accuracy. Without a decoder that can operate on properties of the vehicle itself to check for high quality in the second-order vehicles, we would have to make third-order representations of the second-order representations, and so forth. Particularly if the subject's relationship to the representational vehicle is one in which the vehicle in its entirety is bound up into the nature of the subject, there is no benefit to not taking advantage of the complete information about the vehicle that is potentially available by using the vehicle to represent itself.

I suggest that second-order phenomenal judgments are based in this kind of an adaptation. The representation vehicles underlying judgments such as, "I am having a sensation of purple" and "I am experiencing double vision" are the same representation vehicles used in first-order judgments. However, they are being exploited by different guidance-taking mechanisms, mechanisms designed to take advantage of the potential created by the presence of phenomenal properties in the relation of acquaintance: Full information about them is potentially available to decoding mechanisms. By being able to guide different decoding mechanisms, phenomenal properties come to have different representational content depending

on which decoding mechanism is dominant in a given circumstance of the subject. If these mechanisms are functioning normally, they can deliver certain a posteriori knowledge because, first, there is no metaphysical distance between decoder and object and, second, the object is being used to track itself, so there is no representational distance between representational vehicle and content. Second-order phenomenal judgments represent a failure of diaphanousness, as phenomenal contents are used by alternative mechanisms to represent themselves, not objects external to cognition.

Not all second-order judgments will yield certain knowledge. The decoding mechanism itself will have structural limits that can lead to errors. For example, experiments show without a doubt that these mechanisms (1) have bandwidth limitations, (2) have storage limitations with respect to memory, and (3) are limited with respect to how fine-grained their measurement of similarity and difference can be. Limitations of type (1) show themselves in experiments that demonstrate the difficulty of simultaneously attending to experiences in different sensory channels, visual and auditory, for example. Limitations of type (2) show themselves in experiments demonstrating change-blindness, in which subjects given two successive but slightly different visual scenes will not be able to notice the difference. Limitations of type (3) show themselves in color or sound discrimination experiments in which subjects are unable to judge as different two colors or sounds very close to one another in sensory space. If these processing limits are strained, our decoding mechanisms cannot function properly, and certainty in second-order judgments cannot be achieved. Because ordinary experiencing contains a vast amount of phenomenal information at any moment, far more than we can attend to given bandwidth, storage, and discriminatory limits, ordinary second-order judgments are subject to relatively large amounts of doubt.

Nevertheless, within these processing limitations, second-order phenomenal judgments can deliver certain knowledge. For example, on looking at a tomato, we are able to attend to our color experience and, by attending to it, know that we are having an experience of phenomenal redness and that we are having an experience of color and know these things with *certainty*. We could never have such certainty if the redness was attributed to an external object which is at a representational distance from us, nor even if we had a second-order representation of our first-order representation, for the same reason. Certainty can be delivered only in a properly functioning subject and only by simultaneously closing the metaphysical distance between the metaphysical knower and the known as we close the epistemic distance between representation and representational vehicle. The moral is:

> Our concept of phenomenal redness contains phenomenal redness as its representational vehicle, and we, as subjects, are acquainted with this vehicle. A similar conclusion applies for other phenomenal properties for which we can form distinct concepts.

Finally, with third-order phenomenal judgments, certainty is more common because type judgments strain capacity limitations far less than do judgments about

the moment-to-moment cacophony of temporally passing experience. To make a type judgment, a subject's decoding mechanisms need only to be sensitive to properties of the representational vehicles as they are isolated by attention when they are isolated across a lifetime. The decoding mechanism does not need to be able to stop and isolate a large amount of phenomenal information from many active sensory channels at a time. Instead, the subject needs only to construct judgments about these properties over time, through repeated acts of attention, as attention isolates elements of experience again and again in different instances.

In this picture, the gap between *knowing what* and *knowing that* is crossed via possession of a secondary interpretation mechanism. This interpretation mechanism embodies procedural knowledge, that is, it is a *knowing how*. If this is correct, then it is a *knowing how* that is present innately in a crude form and able to be refined through training that closes the gap between *knowing what* and *knowing that*.

It seems that, if the Consciousness Hypothesis is true, the knowledge paradox regarding consciousness is resolvable. Even though consciousness is not physical, its activity underlies our physical nature as a carrier of our nomic content. Our physical states, although not causally interacting with our conscious states, track and therefore represent those states. The relation between the subject and its experiences is one of acquaintance, in which the metaphysical distance between experiencer and experienced is closed, and within acquaintance the relationship between representational content and representational vehicle can also be closed: Phenomenal properties are used to represent themselves.

The superfluity of consciousness. The problem created by the superfluity of consciousness concerned the challenge that an epiphenomenal consciousness would pose to scientific realism. Because experience is a fundamental carrier and because carriers are required elements in causation, the problem of superfluity is solved directly. This claim can be illuminated another way by considering the Causal Exclusion Argument introduced by Jaegwon Kim (1993, and discussed extensively by him in Kim 2000) and predicated on the situation represented in one of its forms in figure 13.2 (it can also be run with a causal relation from M1 to P2 rather than M1 to M2).

Figure 13.2 depicts a physical state P1 that is supposed to realize a mental state M1 and a physical state P2 that is supposed to realize a mental state M2. Imagine that P1 and P2 are neural states and that M1 and M2 are psychological states. Here, realization is taken to be a necessarily sufficient condition that is not identity, and the property of having a physical constitution P1 (or P2) is supposed to be on the same level of nature as the property of having the mental state M1 (or M2). Also, P1 is supposed to cause P2, whereas M1 is supposed to cause M2 (sometimes the causal relation is drawn from M1 to P2). We can take the causal relation to be a nomologically sufficient condition. The causal exclusion argument is that if P1 is sufficient to cause P2, and if P2 is sufficient to realize M2, then there is no need for the causal relation between M1 and M2. It is redundant.

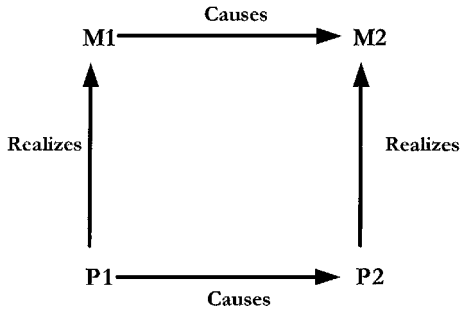


Figure 13.2 A representation of one form of the traditional causal exclusion argument.

Kim uses the Causal Exclusion Argument to argue that nonreductive physicalism reduces to epiphenomenalism (i.e., the causal irrelevance of the mental). More generally, it can be taken as presenting a complex dilemma: Either reductive physicalism is true, interactive dualism is true, epiphenomenalism is true, or downward causation is true.

Figure 13.3 depicts how the situation projects under the Theory of Causal Significance.

The diagram in figure 13.3 violates the assumptions of the traditional diagram in several crucial ways. First, if the Consciousness Hypothesis is true, then the relation connecting the neural physical states to the higher level mental states is no longer one in which the physical states alone present necessarily sufficient conditions for the corresponding mental states. It does so only in conjunction with a specific structure of receptive connections creating layers of natural individuals and ending in a cognitively structured high-level individual. For example, for a neural level state P1 to realize M1, there would need to be several levels of irreducible receptive connections binding the physical elements of P1 into layers of higher level, natural individuals, eventually incorporating a cognitively structured individual. In fact, we would be more accurate if we thought of the relation

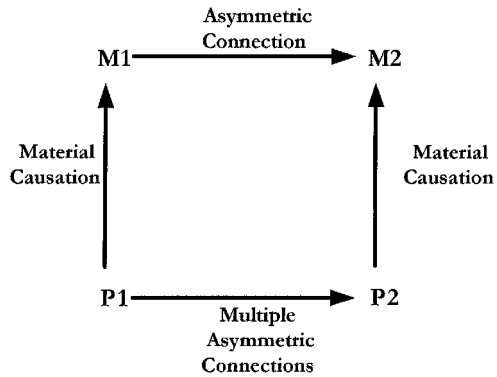


Figure 13.3 The Causal Exclusion Argument if the Theory of Causal Significance is true.

between P1 and M1 (or P2 and M2) on the model of the older concept of material causation rather than realization. Therefore, we cannot properly think of P1 and M1 as lower and higher order properties of the same-level individual.

Furthermore, in each asymmetric connection, one member conditions the other, and the higher level connection is not at all redundant relative to the lower level connection. For example, the joint states of P2's physical elements must be indeterminate when considered independently of its environment and of M2, and in reality there would not be just a single causal connection between P1 and P2 at the level of their basic physical members. More accurately, there would be a multiplicity of connections between individual physical elements of P1 and individual physical elements of P2. Assume this multiplicity of connections and also that P2 has three independently possible joint states for its physical elements. Label these three possible joint states $P2_1$, $P2_2$, and $P2_3$. P1's causal significance should condition away one or more possible joint states in P2, such as joint state $P2_3$, leaving the others, such as $P2_1$ and $P2_2$, as the two remaining possible joint states. This would be an example of efficient causation between the physical elements of P1 and P2 at their own level.

We can imagine that there is only one level separating P1 from M1 (and P2 from M2). This implies that for M2 there are two independently possible states, corresponding to $P2_1$ and $P2_2$. The receptive connection through which M1 may as a whole further condition M2 as a whole should allow it to eliminate one of the states, such as the state corresponding to $P2_2$. The causal significance of M1 thus would leave M2 in a determinate state corresponding to $P2_1$ when it would otherwise be indeterminate. This is efficient causation between M1 and M2 at their level, and it is not redundant relative to the efficient causation at the lower level. Also, the implicit dilemma posed by the causal exclusion argument turns out to be a false dilemma: M1 is clearly not epiphenomenal, nor does it reduce, nor is it interactive, nor is its influence an instance of downward causation. This last claim, that it is not an instance of downward causation, is one that I examine in detail in the next chapter.

The grain problem. First proposed by Wilfrid Sellars (1963a), the grain problem contrasts the homogeneity of experiential qualities with the structural complexity of their supposed physical basis. In chapter 7 I suggested that we might be able to solve the problem by carving off the functional being of mental events from their physical basis, focusing only on their level-encapsulated causal roles within a canonical context. The Carrier Theory of Causation proposed here, combined with the Consciousness Hypothesis, achieves the appropriate isolation of functional being from physical basis.

If the Consciousness Hypothesis is true, then the hierarchy of natural individuals defines the levels of nature, and the causal nexus provides a canonical context within which we can define the functional being of an entity.

Definition 13.1. The functional being of an individual I_n that is a member of a higher level individual I_{n+1} consists in the set of effective contributions that I_n

may make to constraints on the states of other individuals in the canonical context provided by I_{n+1}, as well as its responses to the constraints to which it may be subject, in the variety of causal circumstances within which I_{n+1} may find itself.

Recall that the discussion in chapter 10, section 4, about diagram (g) explained that we can and should identify an individual's effective properties with the elements of constraint it adds to the nexii in which it exists. The motivating observation was that these contributions to the constraints within a nexus are multiply realizable, meaning that two structurally different states of different individuals may map to the same effective property. If two different effective states of two differently structured individuals make the same difference to constraints within a nexus, from the point of view of the other members of the nexus the differences in internal structure are irrelevant and ultimately lost.

Because effective properties are emergent from effective states, it is parsimonious to suppose that there is a single carrier and no more than a single carrier for each effective property. If so, there should be a mapping from each effective property realized by a member individual's effective state to one specific carrier for that effective property. The carrier, then, could not be structured to match the internal structure of a member's state because it would have to stand ready to carry its contribution to the nexus's constraint as presented equally by a variety of different possible members, each of which may be structurally different from the others.

In fact, the most logical structure for carriers to have would be one matching the dimensional structure of the constraint it carried. This dimensionality would constitute its degrees of freedom as a part in a signaling system used by a natural individual's members because the carriers, when viewed this way, are clearly carrying signals that show the character of the constraints active between the bound individuals. These signals could even find reflections in the physical image. For example, our color processing system signals to the rest of the visual system using a three-dimensional vector code that varies along the dimensions of hue, brightness, and saturation. Together, these signals embody the effective influence of color on visual and cognitive processing. Because phenomenal carriers carry the effective influence of these signals, one would predict that the phenomenal carriers would also have a three-dimensional structure.

The boundary problem for experiencing subjects. According to the Central Thesis, an experiential property carries receptivity, and phenomenal properties carry effective properties. It follows immediately that the world's receptive structure, that is, the facts about what is receptively bound to what, determines the boundaries of experiencing subjects. Although only an idealized "in principle" solution to the problem, it meets the challenge laid down earlier: Identify what it is that grounds the possibility of mid-level experiencing subjects bounded in just the way human consciousness is bounded. The answer is the presence of receptive boundaries at many different levels of nature, including the level of brain activity associated with conscious experience.

The remainder of the problem is by far the most difficult part: Adapting the theory of natural individuals to real science to determine the actual structure of the causal mesh. This is clearly a highly nontrivial task, although I make some first suggestions in the next chapter. Nevertheless, the conditions of natural individuality make logical sense of the problem. To be a natural individual is to be a completed receptive connection. As a causal nexus bounded by the carrying capacity of a real connection, each natural individual provides boundaries for the receptive field of its own elements: Only within it may they directly interact by exchanging carriers in the act of completion. Also, the inductive definition of natural individual makes for a very nice way to extend natural individuation to the mid-level of the world, where it is needed if we are to explain the existence and boundaries of consciousness. These are nontrivial objective conditions for inherent individuality that go beyond mere spatiotemporal patterns.

13.4 Making Peace with Receptivity

I initially made the case for receptivity entirely on philosophical grounds. The philosophical argument is roughly this:

1. We need a detailed metaphysical model of causal interaction if we are to understand the problem of inherent individuation in the world (chapter 4).
2. There are also several secondary reasons to want a theory of causal interaction if we are to resolve puzzles and paradoxes surrounding consciousness (chapter 7).
3. A model of causal interaction must be realist as opposed to Humean (chapter 8).
4. A realist model of causal interaction must contain effective properties (chapter 9).
5. Conceptually, a realist model of effective properties implies a realist model of receptive properties (chapter 9).
6. Accepting the duality between these properties provides a model for making sense of the causal powers of various proposed historical scientific and philosophical entities, such as different varieties of space, singularities, epiphenomenal minds, and God (chapter 9).
7. Therefore, here is the simplest model of causal significance that respects this duality (chapters 9 through 11).

When one is first exposed to the idea of receptivity, one's intellectual interest in the idea is sometimes accompanied by real misgivings. Ultimately, the misgivings divide into two categories, conceptual and evidential. On the conceptual side, people worry about the kind of conception we can have of something like receptivity characterized, as it is, as an entity that is a kind of incomplete but pure openness. It can seem very hard to visualize appropriately. On the evidential side, people worry how we could ever get access to the supposed facts about

receptivity. Because the physical facts underdetermine the receptive facts, the story about receptivity may seem like rationalist speculation, unfettered. Gripped by these concerns, people can sharply feel the temptation to explain receptivity away in some fashion. Even if receptive connections seem metaphysically required to account for the causal nature of the world, they seem conceptually obscure and epistemically opaque, and we may react with an urge to explain them away.

Why shouldn't we follow this skeptical urge? In looking for reasons to resist the urge, it helps to gain perspective on what these last few chapters have revealed about receptivity and the character of the receptive connection. We have now deduced a set of empirical predictions from the model and added them to the initial set of philosophical reasons for accepting the existence of receptive connections. The previous discussions have shown that receptivity is explanatorily irrelevant from the point of view of physics yet still necessary for a full explanation of the causal structure of the world; its exact relation to the effective (i.e., physical) properties is sui generis due to a kind of compositional interdependence; the causal nexus created by a shared receptive connection must have some kind of partless unity; the receptive carrier will have a kind of neutral openness to properties that are intrinsic *tout court* (e.g., phenomenal content); receptivity has a sui generis relation to physical spacetime; shared receptivity defines inherent individuality in nature; the carrier structure of the receptive field will reflect a signaling/information structure in the natural world; these inherent individuals and information structures could exist at many levels, including a mid-level; within a cognitively structured individual, binding can provide a kind of knowledge by acquaintance; and one can deduce the privacy of an individual's experience from the privacy of its receptive field.

The impressive similarity between these attributes and traditionally problematic claims about consciousness is eye-catching. These attributes are precisely reminiscent of the kinds of features consciousness typically has been thought to have and the kinds of explanatory problems it presents. I stress—and this is important—that the entire story about the receptive connection and its queer nature was motivated and developed *independently* of the problem of consciousness. On this independent basis, it seems that we can deduce that there is *something* with these strange properties.

Repeating, receptivity seems to be metaphysically required as part of the causal structure of our world, yet it seems conceptually obscure and epistemically opaque. I now point out that consciousness presents itself as epistemically transparent and conceptually immediate: We have observational knowledge of it (chapter 2). Yet it seems metaphysically baroque, so we do not have full confidence in our observations. Not only are the phenomenal qualities that exist within consciousness brute features of nature, but also many of the apparent features of that experiential context seem extravagant, and accounting for them is awkward. The entire package is unmotivated by any deeper naturalistic considerations. Why should the world contain such a thing?

As an intellectual poser, consciousness is the mirror image of receptivity. The problem is not observing that it exists or that it has many of those strange features. Evidentially, it and they are presented to us. But it is very difficult to *believe* what is presented to us because, metaphysically, it seems too queer and unmotivated a kind of thing; it has no natural place in the world. It is just a strange "nomological dangler" on an otherwise internally complete and self-consistent machine, a physical machine belonging wholly to the physical world.

By adopting the Consciousness Hypothesis, the Liberal Naturalist can use the non-mysterious features of each entity to address mysteries arising with regard to the other. A Theory of Natural Individuals incorporating receptivity predicts many of the most troubling aspects of phenomenology on independent grounds. The attractiveness of this strategy is obvious because it is clear that, if we adopt it, each strange phenomenon, consciousness and receptivity, undercuts the motivation for skepticism about the other. The receptive connection is epistemically opaque, but consciousness is not, and so it can be a model for a real, live caught-in-the-trap receptive field. The phenomenal field of consciousness seems too strange to be what it seems, too arbitrary and brute, but the characteristics of the causal nexus created by receptivity are not arbitrary. Consciousness is strange in just the way a carrier of nomic content has to be strange.

This achievement is no mean feat, and it should not be cast aside lightly. We have looked without blinking into the depths of the natural world. In those depths we have found possible truths about the categorical foundations of causation itself, possible truths whose substance, strangeness and importance reflect the exotic depths from which they have come. The Liberal Naturalist urges that it would be both more interesting and more fruitful to accept the existence of both consciousness and receptivity than to yield to skepticism, and so we should carry out the project of developing the resulting view of nature.

Also, the Liberal Naturalist who adopts the Consciousness Hypothesis exposes the false dilemma presented by the causal argument against antiphysicalism. When asked if consciousness is a ghostly Cartesian entity mysteriously interacting with the physical world or an ugly "nomological dangler" irrelevant to it, the Liberal Naturalist may answer: neither. For physicalism to press the issue, it must address the premise that causation is entirely physical in its argument and must do it in a way that goes beyond pointing to the success of the physical sciences. The burden of proof now shifts, as we know that this physicalist argument is unsound under at least one substantive view of causation motivated from first principles, independently of the problem of consciousness and compatible with the success of physical science. What is the physicalist's theory of causation and what are their carriers?

14

Applications

14.1 The Theory of Natural Individuals

The purpose of this book has been to place consciousness in nature. To find a place for consciousness, the previous chapters have developed a framework for understanding causation. This framework is the Theory of Natural Individuals, and, through it, experiencing is tied to the deep structure of the natural world. The three primary elements of the framework are:

- The Theory of Causal Significance
- The Carrier Theory of Causation
- The Consciousness Hypothesis

Having a good framework is critical, but it may require us to undertake substantial further thought before we can fully apply it. In the last chapter, I applied the framework to the several issues raised in part I of the book, but there are many other important areas to which the Theory of Natural Individuals applies.

The Theory of Natural Individuals touches the philosophy of mind, physics, and cognitive science. Detailed discussions of how the Theory of Natural Individuals might be applied in these areas could probably fill several sequels to this book. I make some first comments in this chapter, but I do not pretend to do the issues justice. I only hope these sketches and observations, though inadequate, will help further clarify the commitments and usefulness of the framework. Therefore, this chapter discusses briefly the touch points between the Liberal Naturalist framework that I have called the Theory of Natural Individuals and a set of further applications of the framework.

14.2 Philosophical Applications

I believe that the Theory of Natural Individuals is a synoptic metaphysics that bears on most of the major problems within philosophy. If it holds, it may have consequences for discussions of modality, the nature of concepts, Platonism, skepticism, free will, value, and intentionality, among others. I encourage readers who may be interested in these topics to consider for themselves how the framework might bear on them. In this section I discuss briefly its relevance to three philosophical questions:

1. Are there strongly emergent properties above the level of fundamental physics?
2. What is the precise causal relevance of consciousness?
3. How does the Consciousness Hypothesis bear on functionalism?

14.2.1 Emergence

There are two notions of emergence. The first notion, which I call *weak emergence*, is noncontroversial. It refers to nonfundamental properties such as liquidity, shape, solidity, and flammability that emerge in a constitutive way from the organizations and interactions of lower level entities. In the terminology of chapters 2 and 3, the lower level facts entail the facts about these properties. They "emerge" in the sense that they are numerically different from any lower level properties, but they are not radically novel properties because their instances are explicable as the inevitable consequences of the activity at the lower levels.

The second notion of emergence, which I will call *strong emergence*, refers to the appearance of new fundamental properties that exist only at the higher levels of nature. It is controversial whether any strongly emergent properties exist, and orthodox belief is that they do not. Strongly emergent properties are properties whose instances, if they exist, are not wholly constituted by the organizations and interactions of lower level entities, although their existence may be a consequence of the lower level activity in conjunction with suitable fundamental laws that apply specifically to the situations in which they emerge. One might say that the strongly emergent properties are not *constituted* from lower level activity but are *generated* or *materially caused* by that activity.

The question naturally arises as to whether consciousness is *weakly emergent* or *strongly emergent* or some combination of the two. Unlike physicalism, which is committed in spirit, if not in principle, to the weak emergence of consciousness from physical facts, different Liberal Naturalist theories might say different things about the emergence of consciousness. Depending on the specifics of the Liberal Naturalist theory, consciousness could be weakly emergent from some nonphysical facts (e.g., from instances of protophenomenal properties) or strongly emergent even given the nonphysical facts.

The Consciousness Hypothesis claims that the elements of consciousness are

the intrinsic carriers within cognitively structured, high-level natural individuals.[1] According to the theory, the intrinsic carriers of consciousness come in two fundamentally distinct types—one type that carries receptivity and one type that carries effective properties. The emergence question must be answered separately for each type of carrier. The emergence question applies directly to these intrinsic carriers: Are they weakly or strongly emergent?

Receptivity. The intrinsic carrier for receptivity corresponds (roughly) to the traditional experiencing subject, although it does not have much of the baggage usually associated with the experiencing subject. For example, the receptive carrier does not have the burden of being a persistent self. Rather, it is a connective property whose experiential nature carries receptivity. Within this framework, a natural individual's intrinsic receptivity is a nonreducible connective property binding to all the other individuals in a nexus and belonging to the higher level individual so constituted. By definition, a receptive connection is a neutral essence that is not affected by other instances of receptivity and that is confined to the individual it helps constitute. It follows that the receptive properties belonging to the individuals at each level of nature must be strongly emergent properties, each instance acting as an irreducible global property of the higher level natural individual to which it belongs.

Phenomenal properties. Phenomenal properties carry effective properties within an individual's receptive field. It is less clear-cut whether phenomenal properties are weakly or strongly emergent. For the Carrier Theory of Causation to hold, there must be a function from the carriers of an individual's effective states to the carriers of that individual's effective properties. In principle, this function perhaps could be instantiated by a compositional rule that allows for weak emergence. To handle cases of multiple realization an appropriate weak emergence rule would have to explain how two or more different effective states, carried by two different sets of lower level carriers, each combine to form qualitatively identical carriers for the single effective property realized by them within higher level individuals.

Although weak emergence might be possible, it is also possible that the function from effective states to carriers is realized in a strongly emergent way. If so, the function would be instantiated by a fundamental operator mapping lower-level effective states onto strongly emergent carriers of effective properties, according to the constraint contribution placed by the lower-level states. Although no considerations are decisive, there are some reasons to prefer the view that the phenomenal carriers at each level are strongly emergent.

To understand these reasons, consider figure 14.1. It depicts six level-zero individuals ($I_{0.1}$ through $I_{0.6}$) bound into two groups of three within two level-one individuals ($I_{1.1}$ and $I_{1.2}$), which in turn are bound into a single level-two individual, $I_{2.1}$.

According to the Theory of Causal Significance, each individual has an effective state, which is just the ordering of the effective properties possessed by the

individuals bound within it. In this example, the effective state of individual $I_{1.1}$ is the ordering of whatever effective properties its bound level-zero individuals $I_{0.1}$, $I_{0.2}$, and $I_{0.3}$ have. Similarly the effective state of individual $I_{1.2}$ is the ordering of whatever effective properties its bound individuals $I_{0.4}$, $I_{0.5}$, $I_{0.6}$ have. When the same kind of constraint contribution may be placed by more than one kind of effective state, then the same effective property can be realized by multiple effective states. It follows that an individual's effective states and effective properties are distinct.

According to the Carrier Theory of Causation, within each natural individual there are phenomenal properties carrying the effective properties of its bound individuals. The question now raised is whether the phenomenal carriers within an individual such as $I_{2.1}$ are more likely to be weakly or strongly emergent relative to the phenomenal carriers within $I_{1.1}$ and $I_{1.2}$. Observe that the differences any individual makes within a nexus it joins are *informational* differences. Because effective properties are multiply realizable components of constraints, the physical constitution of an individual is essentially masked within the nexus by the constraint character of its effective properties. The only differences in effective states that get communicated within the nexus are those differences that make a difference to the constraints it places. These *differences that make a difference* constitute an information structure within the nexus, representable abstractly and without reference to details of inner constitution.

In computer science terms, an individual's irreducible receptivity, through which it holistically receives constraints within the nexus, and its effective carri-

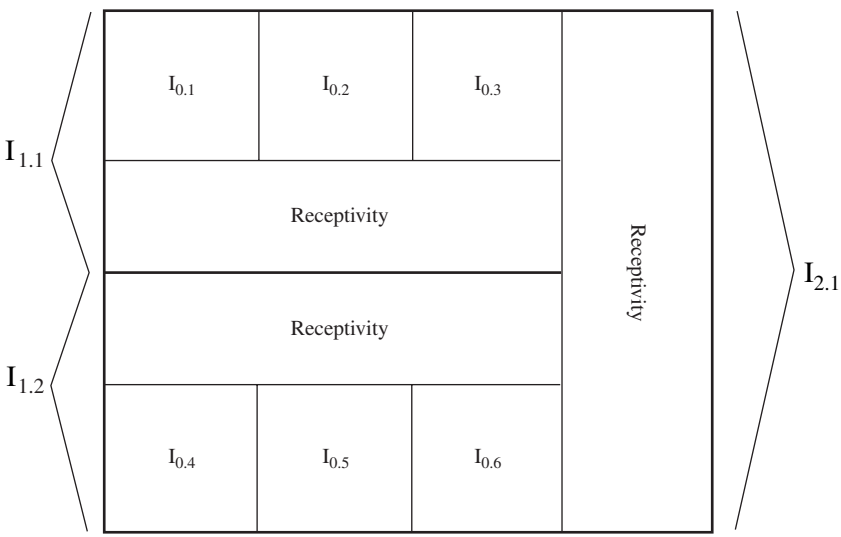


Figure 14.1 A level-two individual. It binds two level-one individuals, each of which binds three level-zero individuals.

ers, through which it places constraints on the nexus, constitute an *interface* between itself and the other individual members of the nexus. The purpose of an interface is to encapsulate the inner nature of an object, allowing it to receive and pass level-specific messages between itself and other objects.

Typically the encapsulated character of an information system is built on a *signaling* system shared with other devices, and the natural way to think of the carriers within a nexus are as signals generated by, or on behalf of, the bound individuals, which one could then think of as signaling devices. Signaling devices typically do not constitute the signals they send in the way a weakly emergent property might be constituted by lower level activity. It is also an interesting fact that natural individuals, no matter what their level, are fundamental individuals. Though they are not necessarily fundamental particles, each of them has an ontological distinctiveness similar in important ways to that of a fundamental particle. When fundamental particles generate signals in spacetime, those signals are themselves such things as fundamental forces carried by virtual particles, where virtual particles are strongly emergent from their particles of origin.

This speculation suggests an analogy: Virtual particles are to internexus signaling between fundamental particles as effective carriers are to intranexus signaling between bound natural individuals. The analogy implies that the effective carriers at each level of reality would be *strongly emergent* rather than weakly emergent, being generated by or on behalf of the natural individuals of that level when they become bound within higher level individuals. If true, this would mean that we should expect the quality space of carriers to be structured into separate irreducible quality families not constituted from one another.

If the strong-emergence hypothesis is true, these families would need to be reachable from one another in some systematic way. The natural laws we should expect to exist are laws that enable nature to take a systematic walk through this quality space, mapping individual effective states onto quality spaces in an orderly way. By analogy, these laws would be something like a mathematical operation, such as vector addition on a set of vectors, which takes a set of vectors as input, and maps to a new vector in the space. It instantiates a many-to-one function that in no way yields a vector "constituted by" the old vectors.

14.2.2 The causal role of consciousness

In the last chapter, I briefly noted that the Consciousness Hypothesis implies that consciousness evades being epiphenomenal by being a carrier for nomic content. It is clear that the relationship between a carrier and a dispositional property it carries is like the relationship between aspirin and the property of being able to ease pain. The carrier is a basis for the disposition and intuitively has the primary causal significance. Therefore, because they are carriers of nomic content, there is no interesting sense in which experiencings or experienced phenomenal content should be considered epiphenomenal. Also, receptive experiencings do not interact with the physical, if "interact" means making an effective difference to

the world's physical dynamics.[2] Therefore, conscious experiencings are neither epiphenomenal nor interactive.

I also briefly discussed the theory's relevance to Kim's causal exclusion argument, claiming at the end that it evaded Kim's four part dilemma by not being reductive, epiphenomenal, interactive, nor entailing downward causation. Yet the previous subsection argued that effective carriers might emerge strongly and carry the effective properties within a nexus. Together, these two suppositions raise questions about whether there is "downward causation" in the theory after all.

Downward causation is the view that there are strongly emergent properties, paradigmatically elements of consciousness, that exert influence on the microphysical behavior of particles, causing them to behave in ways not explainable by their physics alone. In practice, downward causation is much like dualist interactionism, differing from it mainly by placing the interactive conscious properties nearer to the traditional physical world than to a world of substances distinct from the material.

The causal role of the effective properties was most precisely explained in figure 10.12, diagram (f). The text there discussed examples showing how higher level causal nexii act as possibility filters. The natural individuals at each level of nature help solve the determination problem by making the state of the world more determinate relative to the lower levels. In trying to understand better how the role of phenomenal properties compares and contrasts to the role of properties in a prototypical downward causation scenario, I wish to appeal to Aristotle's catalogue of the four causes.[3] Aristotle's views on the four causes give us a finer grained way to distinguish between different types of causal responsibility than do modern ways of looking at causal responsibility. Aristotle distinguished between four different kinds of "causes":

1. *Efficient causes.* The efficient cause of an event or a thing is the primary source of change through time. For example, the cue ball hitting the triangular stack of billiard balls to start a game of pool is the efficient cause of the scattering of balls that follows. Efficient causes lie closest to the modern concept of cause and causal responsibility.
2. *Material causes.* The material cause of a thing is the substance that it subsists in or comes out of. The material cause of a rubber ball is rubber, for example. Under physicalism, physical properties are proposed to be the material causes of everything.
3. *Final causes.* The final cause of a thing is the purpose or end-nature of a thing. The final cause of football players wearing their pads and helmets in a football game is avoidance of injury. The final cause of a sapling is to become a mature tree.
4. *Formal causes.* The formal cause of a thing is whatever it must be by definition, so that it has a property in virtue of being the type of thing it is. For example, the Greeks believed the formal cause of death in human beings is that they are mortal beings.

Notice that Aristotle's four types of causes do not exclude one another. Something may at the same time have an efficient, material, final, and formal cause. For example, a live birth of a baby may have as its efficient cause the insemination of an egg by a sperm and the subsequent gestation; as its material cause the organic molecules of the child and mother; as its final cause the independent life of the child; and as its formal cause the fact that the mother is a mammal (giving live birth is a defining characteristic of mammals).

In downward causation or Cartesian interaction scenarios, conscious events are clearly efficient causes of microevents. They typically are supposed to present some new force or influence on the concrete dynamics of microphysical entities, leading them to change the concrete behavior they otherwise would exhibit. This is not an adequate description of the mechanism by which high level individuals have causal significance. To better understand the causal relations in the new framework, let us create an example and revisit the logic of causal significance in some detail: Consider three individuals, $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$, each independently capable of taking on one of three states representable as –1, 0, or 1. Considering $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$ independently, there are twenty-seven possibilities for their joint state, given by the Cartesian product: $I_{1.1} \times I_{1.2} \times I_{1.3}$, as shown below $< I_{1.1}, I_{1.2}, I_{1.3}>$.

<-1,1,-1> <-1,-1,0> <-1,-1,1><-1,0,-1> <-1,0,0> <-1,0,1> <1,1,-1> <1,0,1> <1,1,1>
<0,-1,-1> <0,-1,0> <0,-1,1> <0,0,-1> <0,0,0> <0,0,1> <0,1,-1> <0,0,1> <0,1,1>
<1,-1,-1> <1,-1,0> <1,-1,1> <1,0,-1> <1,0,0> <1,0,1> <1,1,-1> <1,0,1> <1,1,1>

Assume that a higher level individual $I_{2.1}$ binds $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$ within itself. Also assume that there is a causal law that the effective states of all bound elements within nexii of $I_{2.1's}$ type must sum to 0. Given this assumption, within $I_{2.1}$, the original space of twenty-seven possible joint states for $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$ is shrunk to seven possible joint states, as shown below $< I_{1.1}, I_{1.2}, I_{1.3}>$.

<1,-1,0> <0,-1,1> <1,0,-1> <0,0,0> <-1,0,1> <0,1,-1> <-1,0,1>

By undermining the independence of $I_{1.1}$, $I_{1.2}$, and $I_{1.3}$, $I_{2.1}$ has filtered out twenty of their possible joint states, and in this sense $I_{2.1}$ has helped to make the world more determinate.

However, $I_{2.1}$ itself is still indeterminate between seven possible states. Imagine now that $I_{2.1}$ becomes bound within a yet higher level natural individual $I_{3.1}$. Also bound within $I_{3.1}$ is $I_{2.2}$, another natural individual of the same level as $I_{2.1}$. Considered independently of $I_{3.1}$, suppose $I_{2.2}$ may take on one of two states, whose constituent structure we ignore and just call states *A* and *B*. Also assume that $I_{2.2}$'s state *A* is not compatible with any of $I_{2.1}$'s potential states and that $I_{2.2}$'s state *B* is compatible only with $I_{2.1}$'s potential state <0,0,0>. Under these circumstances, their binding within $I_{3.1}$ would make $I_{2.1}$ and $I_{2.2}$ fully determinate: $I_{2.1}$ would have to be in state <0,0,0>, $I_{2.2}$ would have to be in state *B*, and $I_{3.1}$ would be in an effective state consisting of $I_{2.2}$ and $I_{2.1}$'s joint state. The situation is depicted in figure 14.2.

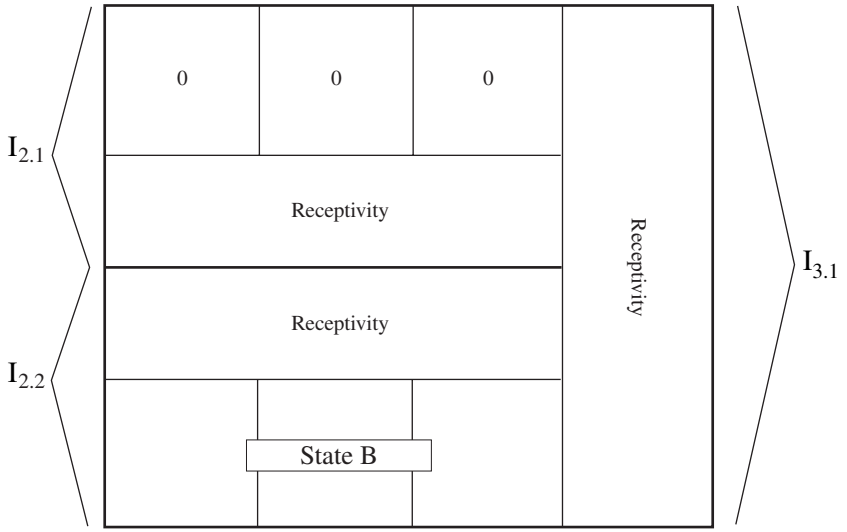


Figure 14.2 A higher level individual, $I_{3.1}$, whose presence makes determinate the lower level individuals within it.

In the example, $I_{3.1}$ represents an individual whose constraint structure is strong enough to produce complete determinacy, and its determinateness implicitly requires completely determinate states for the individuals bound within it. This reveals the sense in which the effective properties of individuals at each level make a causal difference: They influence the possibility space for the joint states of the individuals in a nexus by helping to exclude some of those possibilities.

Please observe three things about this process:

1. The circumstance of a higher level individual coming into being cannot even occur unless the individuals at the lower level are in an indeterminate joint state when considered independently of it.
2. It is inconsistent with the theory for individuals at two different levels to directly interact. The existence of a higher level individual facilitates direct interactions only between its bound constituents, which must be individuals at the immediately lower level.
3. The input space for a higher level individual consists only of the joint states for its constituents that are independently possible given their lower level interactions. The interactions in the new nexus only filter this space, adding no new possible joint states. Therefore, it is logically impossible for the addition of a higher level individual to result in lower level individuals being in a joint state *not otherwise allowed by their own interactions*. It is therefore impossible for the emergence of effective properties within a higher level individual to produce lower level behavior not explainable by reference to lower level interactions (plus randomness).

Essentially, the constraint structure of a nexus is like a set of simultaneous equations, and all possible solutions for the equation are already constrained to be compatible with the results of lower level interactions because those constraints have already been factored into the solution space over which the set of equations is defined. The variables in the equations correspond to the effective properties of the individuals bound within the nexus. The potential determinate states of the effective properties are like the possible solutions for the corresponding variable. Each determinate phenomenal feel, such as an occurrent itch, is an intranexus signal expressing a determinate solution for one variable in the constraint problem. Its presence constitutes part of the effective state of the higher level individual, and it acts as a mechanism for expressing the compatibility between the state of the bound individual whose effective property it carries and the effective states of other bound individuals.

In typical downward causation scenarios, as in Cartesian interactions, conscious events directly interact with lower level entities, influencing their dynamics in ways that violate the behaviors their own interactions would explain. Phenomenal carriers do not really fit this stereotype because higher level individuals operate only on independently allowed possibilities. It is simply not possible for a higher level individual's presence to result in a joint state for its members that violates the lower level physics. What does happen is that phenomenal properties collectively facilitate the immediate global interaction between individuals bound within a causal nexus, individuals all at the same level of nature. At the level of the interactions, the bound individuals are the efficient causes of the nexus's global state.

If the result of this efficient causation is a determinate state for the high-level individual, then the determinate state of the high-level individual may imply that lower level individuals are in determinate states when they otherwise might not be.[4] The implication holds because the lower level individuals are the material causes of the higher level individual, not because there has been an interaction between the phenomenal properties and lower level individuals. The explanation that best fits the situation is one in which the determinate state of the higher level individual acts as a *final cause* or *telos* for the otherwise indeterminate lower level individuals, which have the role of being material causes, and the phenomenal properties play a role in establishing this *telos*.

This is not classical Aristotelian temporal teleology, but it is a kind of teleology nonetheless. It is an atemporal teleology between the levels of nature, one in which the actualization of the determinate higher level individual is the final cause for the determination of the lower level individuals. Note that a kind of anthropomorphic argument applies from the higher to the lower levels. Given that one knows the determinate state of a higher level individual, it makes sense to ask, In what states would the lower level individuals have to be in order to find that the higher level individual is in this state? Anthropomorphic-type questions such as this may imply that the lower level individuals have to be in determinate states themselves, or they may imply that they could still be indeterminate yet restrict the range of indeterminacy. In point of fact, this is exactly what we have

found is true of our world: Sometimes the determinate state of a higher level individual implies that the lower level individuals are determinate, and sometimes it implies that they are indeterminate within a restricted range of values. Experiments confirm that lower level individuals in fact are determinate or indeterminate to just the degree implied by the states of higher level individuals to whom they are coupled. In fact, one possible interpretation of the Theory of Natural Individuals might even be that the lower level individuals don't "really" become determinate so much as higher level individuals must *observe* them to be in states consistent with their own determinate state, because it is only these potential states that make it through the possibility filters. It *may* be that what nature really enforces is consistency between points of view as determined by the structure of natural individuals.

Thus there is a causal role for phenomenal properties as players in efficient causation at their own level and in the final causation of lower level individuals. However, because there is no efficient causation between levels, this situation is not one of Cartesian interactionism or downward causation as those views are usually understood. Figure 14.3 depicts the variety of causal relations supported by this view.

14.2.3 Functionalism

For the last thirty years *functionalism* has been the dominant position in the philosophy of mind. It is certainly the dominant view among practitioners of artificial intelligence and among cognitive scientists. Functionalism is the view that an entity's functional role within the mind is what makes it the kind of mental entity it is. For example, according to functionalism, what makes something a belief is a certain kind of role it plays in enabling rational inference, planning, and reaction, or what makes something a desire is the kind of role it plays in motivation.

The Consciousness Hypothesis clearly implies that *analytic functionalism* about consciousness is false. Analytic functionalism holds that facts about a system's functional organization entail the facts about consciousness because the consciousness facts are analyzable in a definitional way into functional facts. Analytic functionalism is the kind of position that the antiphysicalist arguments discussed in chapters 2 and 3 targeted.

Similarly, the Theory of Natural Individuals shows how *empirical functionalism* also can be false without implying the kinds of consequences often feared. Empirical functionalism is the view that consciousness metaphysically supervenes on functional organization, even though it is not entailed by facts about functional organization. It is in the same family of views as a posteriori physicalism. These were the views discussed in chapter 3, which proposed that consciousness ontologically supervened[5] on the physical despite not being entailed by the physical facts.

There is, however, a third kind of functionalism, introduced by David Chalmers (1996), that I name *nonreductive functionalism*. On nonreductive functionalism, physicalism is false and facts about consciousness are fundamental

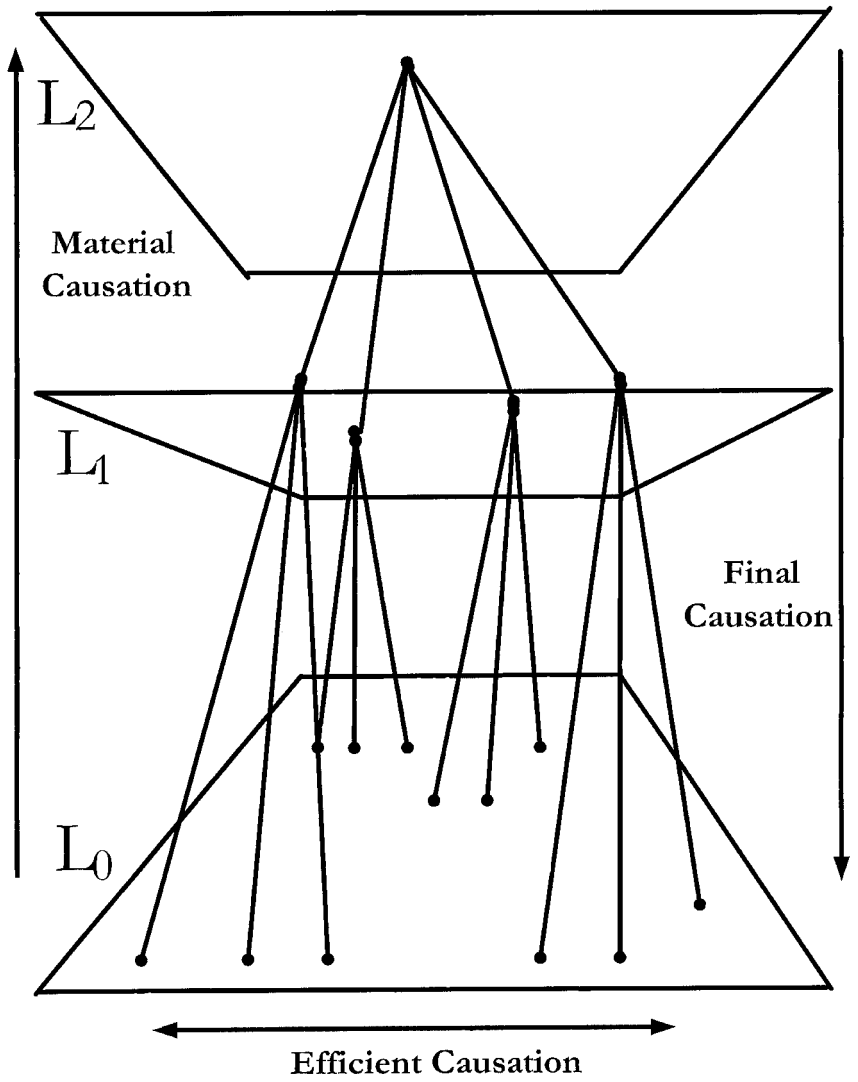


Figure 14.3 Causation within and between levels of nature.

facts that nevertheless are invariant with respect to functional facts. Like those who argue for other versions of functionalism, Chalmers argues that certain highly implausible consequences follow from denying nonreductive functionalism; therefore, we should expect it to be true.

The Consciousness Hypothesis is consistent with a narrowly construed kind of nonreductive functionalism. The narrow construal depends on analyzing a system's functional organization in terms of its causal organization, and its causal

organization in terms of the information structure of the highest level natural individual within the system. The lower level individuals it immediately binds will determine its information structure, and its information structure will be its *intrinsic* causal organization. Two individuals will have the same functional organization if, and only if, their intrinsic causal organizations are isomorphic.

This concept of functional organization based on intrinsic causal organization provides a tighter constraint on the concept of functional organization than one usually finds. It completely eliminates any debate about the observer-relative character of functional organization and eliminates complications arising from the role that counterfactuals play in determining sameness of functional organization. The Consciousness Hypothesis clearly implies that two systems with the same functional organization in this sense will both be conscious and in the same way. Therefore, understood in this tightly defined way, functional organization supervenes on causal organization, the Consciousness Hypothesis holds that consciousness supervenes on causal organization in a parallel way, and Chalmers's functional invariance thesis is not violated.

This view of functional organization has teeth. For instance, it is extremely implausible on this view that neural replacement therapy would change a person's functional organization. The imagined scenarios always imply that exactly the same structure of physical signaling devices is in place, with exactly the same communication pathways between them, that they imply the same messaging systems, and that the same information is being passed between them. Although it is still unclear under just what conditions natural individuals would come to exist, if they exist, it is a good bet that the purely material changes involved in neural replacement therapy scenarios are not relevant. Both Chalmers and Robert Kirk (1994) motivate their endorsements of (their different brands of) functionalism using neural replacement thought experiments, and it is unlikely that the Consciousness Hypothesis yields a kind of functionalism at odds with worries raised by those kinds of scenarios.

It is less clear whether we could download our individual consciousnesses to a mainframe computer. Without a truly detailed understanding of the physical conditions under which natural individuals emerge, we cannot know whether we should expect natural individuals to survive under all transformations of an entity's information processing into another physical embodiment, such as a von Neumann architected computer, just because it preserves causal organization at some formal level. Even at this preliminary stage, it is clear that the existence of a cognitively structured natural individual is going to be tied tightly to physical relations of interaction and communication between a layer of lower level natural individuals. It would not be surprising to find that relations of organization and constitution within space and time matter in a way that rules out an individual's survival through purely abstract mappings of functional structure from one embodiment to another. And it would perhaps be only a little surprising to find out that such spatiotemporal facts make no difference after all. At this early juncture, it is simply an open question.

14.3 Physical Applications

A theory of causation cannot ignore physics. A driving motivation in my construction of the Theory of Natural Individuals was to move beyond Newtonian ways of thinking about causation. In doing that, it seemed to me that the fundamental problem with Newtonian thought was not with Newtonian-versus-quantum or Newtonian-versus-relativistic assumptions but with the implicit assumption that causation is something physically specific to our world rather than something metaphysically general that finds expression in a certain form within our world. An appropriate model of causation would apply to Newtonian physics, to quantum physics, and to all kinds of cellular automata that we could create as models of possible physics in other possible worlds. In computer science terms, I intended to create a *base class* for causation, which could be extended to fit the needs of any particular causal world.

In chapter 9, I introduced and explained the theory of causal significance using some Newtonian examples. But our world is not a Newtonian world, and an adequacy test for the success of the project is the ability to extend the base class to cover the physics of the actual world. Through several of its features, I think we can see that the Theory of Natural Individuals may succeed in specifying such a base class for causation. In several important ways, it seems to evoke a reassuring consistency with quantum mechanics and relativity and makes some of physic's counterintuitive features seem natural:

- The theory's use of determinable effective properties and the role of receptivity in filtering possibilities on joint states evokes superposition and the measurement problem. It actually predicts that in worlds where higher-level individuals exist, the lower-level individuals will be in indeterminate states unless appropriately coupled to higher-level individuals. One can imagine the difficulty a cognitively structured higher-level individual would have isolating those "appropriate conditions." It would look a lot like the measurement problem in quantum mechanics.
- The nature of causal significance is consistent with the existence of quantum entanglement and coherence.
- The view makes nonlocal causation seem expected, rather than mysterious or unexpected.
- Nothing in the theory makes irreducible randomness a surprising feature of the world.
- The spacetime that we could perhaps construct from causal connections, as suggested in chapter 10, begins from relativistic assumptions.
- The potential for nature to have many layers of natural individuals is consistent with the ways that quantum mechanics dilutes the special ontological importance of the microphysical (as pointed out by Lockwood).
- The theory's realism about possibility is consistent with the ability of counterfactual truths to have measurable effects in the quantum world.

In all these ways, one can look at the Theory of Natural Individuals and comfortably say, *if* causation works the way the Theory of Natural Individuals says, *then* it is not so surprising that our physics looks the way it does. In my opinion, it makes quantum physics and relativity seem more natural than classical physics as a way for the world to be.

14.3.1 Physical indications of natural individuality

In the Theory of Natural Individuals, a natural individual is a completed receptive connection. A completed receptive connection has (1) at least one constituent with an indeterminate state when considered independently of its membership in the nexus and (2) a common receptivity being shared by two or more constituents. The shared receptivity establishes a connection between the members of the nexus through which they contribute to a set of simultaneous constraints on their joint states.

Associated with each natural individual I_k is some set of rules, label it Λ, such that, the set of constraints in Λ is most naturally thought of as containing a set of simultaneous equations governing the joint states of I_k's constituents. For each potential effective property of each member of I_k, Λ contains either a variable[6] or a constant[7] referencing that effective property. In a symmetric nexus, Λ will contain variables referencing the effective properties of each member of the nexus. In an asymmetric nexus, only the effective properties of the receiving member(s) will be represented by variables. Individuals whose states are fixed in the asymmetric nexus will have effective properties represented by constant values. Λ expresses a set of constraints on the joint states of the constituents of I_k in terms of those variables and constants. Each solution for the equations represents an independently possible state for I_k.

Additionally, we hypothesize that the bound members within I_k are encapsulated within interfaces. Their interfaces consist of their own receptivities, through which they holistically receive the constraints in their receptive fields, and their own signals, their effective properties, through which they place elements in the constraint structure on the total state of the nexus. These interfaces create an information structure within the nexus.

The above total characterization of a natural individual is fairly substantial from a naturalistic point of view, and it provides some guidance regarding physical world indicators that might be evidence of natural individuality. When searching for natural individuals, this characterization suggests that we should view systems in the physical world as systems of information. For something to be a good candidate for natural individuality, the information system should meet the following conditions:

1. *Base case*: It should be clearly fundamental like a basic particle
2. *Inductive case*:
   - It should be divisible into constituents that are natural individuals themselves, and at least one of which is not known to be in a determinate state, considered independently from the system.

- Its constituent structure should instantiate a system of information-based constraints satisfying the conditions on Λ.
- If thought to be an asymmetric nexus, the receiving member should not be known to be in a determinate state, considered independently of the system, and it should be constrained by the nonreceiving member in such a way as to satisfy the conditions on Λ.

Not all physical systems will meet these criteria. For example, no system consisting entirely of human beings will meet the standard because human beings (presumably) are known (phenomenologically) to be determinate systems when considered independently of other systems of which they might become parts. Also, thermostats and rocks will not be natural individuals because they cannot plausibly be divided into constituents, each of whom is a natural individual and which instantiate a system of constraints meeting the conditions on Λ.

Only three kinds of complex systems will be plausible candidates. Quantum coherent systems would be the first kind, as these systems clearly have globally determined joint states. Other candidates would be synchronous systems built up from rich feedback mechanisms between their constituents. When rich feedback mechanisms exist, we can plausibly hold open that there are holistic constraints on the joint state of a system's constituents that give rise to a constraint structure satisfying the conditions on Λ. Finally, a third kind of candidate system would be temporal processes consisting of natural individuals at each phase, as these will plausibly be chains of asymmetrically connected natural individuals.

What we know about the brain is consistent with these generic criteria. At the brain's coarse-grained level, the communication paths between neural clusters contain high-bandwidth feedback loops. At finer grained levels, neurons within neural clusters participate in feedback systems such as the on-center-off-surround wiring found in the retina. Individual neurons, as cells, are built on rich feedback mechanisms between ionic levels surrounding cell walls and transport mechanisms surrounding energy and genetic material within the cell.

Looking at the big picture regarding the brain's structure, we see that, as a biological system, the brain is hierarchically constructed at many levels from rich biological feedback mechanisms. These feedback mechanisms exist for all physical granularities and are used by the state determination mechanisms associated with individual components from molecules to large neural clusters to entire brain circuits linking together large regions of the brain. Figure 14.4 (taken from James Newman's *Thalamocortical Foundations of Conscious Experience,* 1997) shows the direct inter- and intralevel feedback mechanisms of the circuitry underlying a likely candidate for the neural correlates of consciousness. It shows the presence of both positive and negative feedback mechanisms between neural layers and also between neurons within layers.

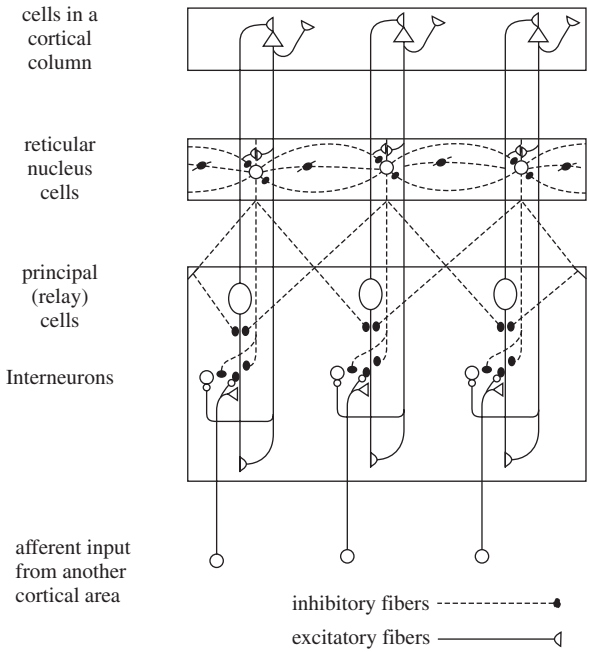


Figure 14.4 Thalamocortical circuit.

The brain also meets the additional criteria that the joint states of its components are not known to be determinate, considered independently of the (proposed) high-level individual. Even though the brain likely does not support quantum coherence, Henry Stapp (2000) argues that, even without quantum coherence, ordinary quantum effects are magnified into state indeterminacy in the brain. His argument is important enough to quote in full:

> [L]et me grant that the chemical interactions could be mocked up by some essentially classical-type model, and that the whole brain can be treated classically except for one thing: the migration of calcium ions within nerve terminals from the exits of micro-channels to the sites where they trigger the release into the synaptic cleft the contents of vesicles of neurotransmitter. The diameters of the micro-channels in cerebral nerve terminals are approximately one nm. This means that the indeterminacy in the velocity of the migrating calcium ion that arises from the Heisenberg uncertainty principle is smaller than its thermal velocity by a factor of about 300.
>
> The distance between microchannel exit and trigger site is about 50nm. Thus the uncertainty in the location of the calcium ion when it reaches the trigger site is of the order of size of the calcium ion itself. This means that the classical conception of the brain is inadequate in principle: quantum effects will generate a superposition of the classical state in which the neurotransmitter in the vesicle is released and the classical state in which this packet of neurotransmitter is not released.

> This superposition will quickly be reduced to a mixture. A similar bifurcation occurs at each active nerve terminal. Hence the state *S(t)_b* will necessarily evolve into a mixture of a huge number of states. Actually, a continuum of possible states will contribute, because each vesicle could be released a little earlier or a little later, and this will produce a continuum of contributing possibilities.
>
> This first step is already important, because it shows that the idea that classical physics could give a deterministic answer to how the brain evolves is in principle wrong: that possibility is strictly incompatible with quantum theory, even if one ignores quantum effects associated with chemistry. Quantum effects entail that the brain state *S(t)_b* will quickly evolve onto a mixture of quasi-classical possibilities, all of which are actually present, insofar as no actual collapse has occurred. It is important in what follows that the interaction with the environment, although it reduces superpositions to mixtures, does not reduce the mixture of quasi classical possibilities to a single one of these possibilities: all states of the mixture will continue to exist in parallel, insofar as the evolution is controlled by the Schroedinger equation.

Stapp is arguing that quantum uncertainties arising in the molecular interactions at neural transmitter junctions create quantum indeterminacies about the facts of transmitter release and reception. For example, quantum uncertainty about the position of a calcium ion can create indeterminacy about whether a neurotransmitter is at a receptor, a release vesicle, or in transit. Neural activation is a function of transmitter release and reception, so these indeterminacies give rise to quantum indeterminacies in the macrofacts about brainwide neural activation. This is just indeterminacy in the joint states of neurons, which is a precondition for the existence of a higher level individual. Importantly, Stapp is arguing that this is true even if we take decoherence effects into account (in fact, he assumes decoherence).[8]

If Stapp is right, the brain is not known to be in a determinate state independently of the existence of a high-level individual. Between this fact and the facts about the brain's hierarchical structure of feedback-regulated components, the brain should be considered a good candidate for supporting the existence of a high-level, cognitively structured natural individual. Therefore, what we know about the brain is consistent with the Consciousness Hypothesis.

14.3.2 The flow of time

In section 10.6, I made some speculative suggestions regarding a method for reducing space and time to relations of causal significance. The reduction proceeded in two steps. In step 1, the method singles out causal processes called cascades, using them as extended frames of reference for creating temporal partitions. Cascades are chains of asymmetrically connected individuals, and each individual link in the chain can be used as a pivot point for creating a temporal partition. The procedure decides membership in a partition by calculating a quantity, called causal distance, possessed both by individuals not in the causal process and by the pivot individual and calculated relative to subsequent indi-

viduals in the process. Any individual that is causally equidistant with the pivot individual to a subsequent member of the pivot individual's cascade belongs to the same temporal partition as the pivot individual. The method yields one temporal partition for each member of the cascade, a set of individuals in the partition, and an ordering between the partitions that mirrors the ordering of the individuals within the cascade. In step 2 the procedure introduced the idea of a signaling path through the causal mesh to reduce spatial relations to specific facts about direct and indirect causal significance.

This view turns intuition on its head. Intuitively, we think of asymmetric causal relations as subsisting in the fundamental asymmetry of time. A reduction of time to causation is a view where the causal relations are fundamental and the temporal relations are logical constructions. On such a view, one naturally wonders where the apparent flow of time fits. On the surface, at least, the construction of time from causation looks relativistic in spirit, and like relativity it seems to involve a "timeless" view of time in which its flow is somehow an illusion. This is a strange consequence in relativity, but there we can at least delay dealing with the strangeness by saying that the flow of time is in consciousness only rather than objectively in the physical world. Here, though, our theory of causation is built fundamentally on panexperiential facts, and the physical world in a certain sense subsists in experiencings. It would seem an odd dodge for a Liberal Naturalist championing the reduction of space and time to relations of causal significance to say that the flow of time is in experience but not in the physical world. Therefore, the problem of the flow of time becomes pressing.

Thus pressed for time, I think Liberal Naturalists have to admit that the proposed reduction of time to causation is incomplete. Specifically, the flow of time is *subjective,* and each causal process consists of linked *subjects of experience*: These are the things acting as the frames of reference within the theory. It is not plausible for the Liberal Naturalist to admit the reality of the subjective and also deny the reality of the flow of time within nature's frames of reference.

However, given that the flow of time is for a subject and that subjects play their roles within frames of reference (causal processes), the Liberal Naturalist may reasonably propose that the reduction of time is accomplished only for *intersubjective time* (I-time). I-time is the ordering and organization of spatiotemporal facts *between* frames of reference. According to this view, I-facts are not fundamental: There are no fundamental facts about space and time between frames of reference. Rather, such facts are constructions derived from more fundamental facts about relations of causal significance. I-time does not flow because it is a logical construction rather than a real dimension to nature.

However, the proposed reduction says nothing about the fundamental reality of subjective time (S-time). S-time is the ordering of temporal relations purely *within* a causal process usable as a frame of reference for the construction of I-time. S-time may perfectly well be fundamental and may flow. In fact, supposing the existence of S-time not only makes sense of our conscious experience in a nonillusory way but also is an attractive mechanism by which nature could

carry the asymmetry of the causal process. That is, the flow of time in experience is a carrier for the structural asymmetry required for asymmetric causal connections. Thus, in contrast to I-time, S-time is real and fundamental and not a logical construction, and its flow has an objective purpose.

A world in which the flow of S-time is real and fundamental and I-time is a logical construction would be partially analogous to the kind of multiverse I described when creating the unity problem for Humean causation in chapter 8. Here is the way I set up that problem:

> [C]onsider a collection of causally separated dimensions, such as a set of parallel universes in a science fiction novel. We can coherently conceive of each separated world as possessing its own internal time dimension. In this kind of multiverse, each world's time dimension would sequence the events within it. Nevertheless, there would not need to be an overarching transworld time sequencing events across worlds. Thus there would be no answer to questions about whether event X in world A occurred before or after event Y in world B.
>
> In the limit, each world could contain only a single event, with an internal time dimension giving it duration. But there would still be no transworld time that ordered events with respect to one another across worlds. A Humean world, with its insulated events, could very well reduce to this kind of multiverse of small worlds: Each event instantiates an internal time dimension that gives it duration, but there does not need to be a common temporal framework within which they all exist. Nothing would order them relative to one another, so they need not form one world rather than many separated, single-event worlds.

A real and fundamental S-time possessed by each causal process is like the "internal time dimension" I ascribed to each of the separated worlds. However, unlike those hypothetical worlds, causal processes are not radically separated from one another. Although they share no overarching S-timelike temporal framework, they do share fundamental internal connections to one another through relations of causal significance, and these connections can be used to construct an ordered I-time. Here, I-time provides a framework modeled on S-time and capable of organizing their different S-time events relative to one another. Yet, even though it provides organization geometrically analogous to S-time, it does not provide flow.

14.4 Cognitive Science Applications

The Theory of Natural Individuals is a framework connected specifically to experiencing by the Carrier Theory of Causation and to consciousness by the Consciousness Hypothesis. Consciousness is a natural phenomenon whose physical aspects are studied directly by cognitive neuroscience. Just as the framework needs to cohere with basic physics, it also needs to cohere with our emerging understanding of the physical activity underlying our conscious states. We are making rapid progress in our understanding of the physical aspects of consciousness, and I discuss two ways the Theory of Natural Individuals seems to cohere with

our emerging understanding of how brain activity corresponds to conscious states.

14.4.1 The structure of phenomenal properties

Consciousness contains an enormous variety of phenomenal qualities. The variety outruns our language, and the subtlety and complexity of structure possessed by the different qualities goes well beyond our ability to self-examine the contents of our consciousness. Even the prototypical qualities of our external senses possess structure that evades straightforward classification. Psychophysics has made clean analyses only of the structure of color and sound. Color can be analyzed in terms of hue, saturation, and brightness, and sound can be analyzed in terms of loudness, pitch, location, and timbre.

Beyond color and sound, things become complicated quickly. There are at least five qualities of taste, which are sweet, sour, salty, bitter, and umami. These classifications are more like "prototypical" categories than clean ones. The taste receptors on our tongues are actually sensitive to the elements of multiple categories simultaneously, and the representation of a taste at higher processing levels seems to be coded across multiple modalities (including texture and olfactory sensors) in a way that adds great complexity to the structure of our taste space, making it hard to map cleanly. Smell is even more complicated because there are hundreds of basic olfactory receptors in the human nose (thousands in the noses of many animals) and because higher level processing also uses cross-neural coding. Differences between the receptors and higher processing in different individuals suggest that, in fact, smell and taste sensations may vary significantly from person to person.

Despite the difficulty in discovering the structure of most phenomenal properties, the Carrier Theory of Causation implies that ultimately they do all possess structure the way that colors and sounds possess structure. The signaling role of phenomenal properties guarantees that they will be structured, because signals are codings. The structure of a code is dictated by the sensitivity of the receiver: The code will vary along just those dimensions and in just those ways that are relevant to potential receivers. Therefore, every coding needs to have a determinate dimensional structure like that had by colors. Colors are three-dimensional phenomena capable of blending with one another and with other phenomenal properties to form complexes consisting of a many-layered phenomenal presentation of properties.

Recall that, within a causal nexus, each bound individual has its own receptivity. If it is in a position to be constrained within the nexus, its receptive field consists of the other members of the nexus. According to the Carrier Theory, phenomenal properties carry the effective properties of these other members, acting as signals generated by the effective states of the individuals within the nexus. As a signal, a determinate phenomenal property represents a single value among the

many possible values for the effective property it carries. As codings, the determinables they instance must have a dimensional structure whose size matches the space of possible values of the effective property, and each of its determinate values must map *systematically* in a one-to-one fashion to a value for the effective property. This implies that a phenomenal carrier will have a dimensional structure matching the dimensional structure of the effective property it carries.

This language of receptive fields—signaling, coding, dimensionality, and coding spaces—corresponds very well with the facts of neuroscience and with the search for *neural correlates of consciousness* (NCC). Neurons and neural clusters likewise have receptive fields defined by the signals they are sensitive to, and the outputs of neurons and neural clusters are naturally represented as coded signals occupying points within a coding space. This is standard scientific practice. Indeed, in the case of visual qualia we have found that the three-dimensional structure of colors corresponds to a three-dimensional coded vector produced by visual processing.

Assume that these color vectors are playing an effective role in determining the joint states of the individuals bound within consciousness and that the value of each element in the vector makes some difference. This assumption is plausible because these vectors are transmitted to the thalamus, which binds them in the cortex with further representations that we have good reason to believe become conscious. The Carrier Theory of Causation implies that there should be a structural isomorphism between neural codings of color and the phenomenal structure of color. It in fact implies that a similar result should hold universally between all kinds of phenomenal properties and the vector codings associated with them. As we learn more about the NCC and are able to isolate various kinds of neural coding activity that underlie the appearance of different kinds of phenomenal properties, the Carrier Theory of Causation will imply that we are also discovering the internal structure of phenomenal properties, whether or not this structure can be isolated by introspection.

Here we see that the Theory of Natural Individuals connects to and supports standard practices and results in neuroscience, despite originating from more metaphysical foundations. Requiring such coherence is a healthy check on the speculative character of the Theory of Natural Individuals, and it provides deeper theoretical support for the assumptions underlying scientific practice in the search for neural correlates of consciousness.

14.4.2 Extended reticular-thalamic activation system

Neuroscience has made remarkable progress in creating an overarching framework for understanding the neural correlates of consciousness. Fascinating and important work by James Newman (1997), Rudolf Llinas (Llinas et al. 1994; Llinas and Pare 1996), Nikos Logothetis (Logothetis and Schall 1989; Leopold and Logothetis 1996), Gerald Edelman (1989), Joe Bogen (1995), Francis Crick and Christof Koch (1990, 1995), Jeffrey Gray (1996), and many others is rapidly

bringing together a high-level anatomical and functional understanding of the brain activity underlying the creation of conscious states. The centerpiece of this understanding is called the Extended Reticular-Thalamic Activation System (ERTAS).

The functional problem addressed by ERTAS is the *binding problem*. The binding problem has two parts. First, how does the brain synchronize all the different representational elements of a conscious state into a unified representation of the self as it is situated in its environment? Second, how does the brain choose to raise one of its several possible competing representations into prominence as the accepted representation in consciousness? The ERTAS view addresses the first part of the binding problem very well and provides a foundation for the second part of the problem. ERTAS is built on the striking anatomical fact that the thalamus, sitting in the center of the brain, seems positioned to act as a gateway between the cortex, where the contents of conscious states seem to be determined, and the rest of the brain, where information about the environment, intentions, and expectations seem to be gathered.

The thalamus is an egg-shaped ball of densely packed neurons with both incoming and outgoing connections to virtually every other part of the brain, including the different levels of the cortex, as well as the brain stem and sensory input systems. The thalamus additionally is partially covered by the reticular nucleus (RN), which is a weave of neurons connecting densely to one another and also having incoming and outgoing connections to cortex. Newman (1997) has said that if the thalamus is the gateway to consciousness, the reticular nucleus is the gatekeeper (see figure 14.5, taken from LaBerge, 1995, p. 161).

Functionally, the thalamus, using the RN, is the chief source of extrinsic activation for the cortex, and it has the striking property of being able to filter potential input to the cortex based on the input's harmony with signals already active in the cortex, where "harmony" is measured by the wave characteristics of the different signals. Crick and Koch (1990) discovered that the two-way circuits between visual cortex and the thalamus/RN are able to produce synchronized neural firings in the 40-70 Hz range, seemingly binding different visual contents in a functional way. More strikingly, Rudolfo Llinas (1994, 1996) reports that waves of these coherent oscillations sweep the *entire cortex* every 12-13 msec and seem to be coordinated by the feedback loops between cortical layers and the thalamus. Given that the cortex seems to contain representational contents from every sense modality and from higher thought centers, Llinas et al. (1994) speculate that these sweeping waves of coherent activity create a reality emulating representation in the central nervous system. It is also striking to note that the thalamus is more densely connected to the cortex than to other parts of the brain and that the signals coming into it from the cortex can inhibit signals from sensory and motor systems. This looks promisingly like an anatomical basis for the top-down processing cognitive scientists have discovered is active in the creation of perceptual representations. See figure 14.6.

Of course, the binding of sensory contents is not the full story about the cre-

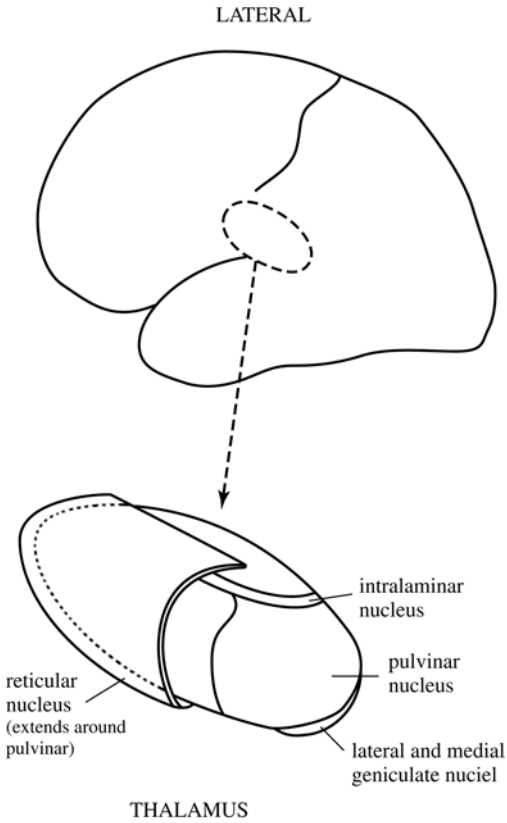


Figure 14.5 The anatomy of the thalamus.

ation of conscious states. For example, Jeffrey Gray (1996) produces data that suggest that the hippocampus also communicates with the cortex and provides a means by which current information can be compared with expectations. But the general mechanisms and the architecture of ERTAS are good candidates for being the anatomical centerpieces from which the brain produces conscious states. Baars (1997), in his commentary on Newman (1997), notes that bilateral damage to the intralaminar nuclei, small eraser-sized pieces of the thalamus on either side of the brain's midline, seems to completely destroy consciousness, creating a very unusual sensitivity replicated only by damage to the brain stem.

From the perspective of the Theory of Natural Individuals, the existence and importance of the ERTAS to conscious states is a hopeful sign. It seems to be just the kind of mechanism we would expect to underlie the creation of a higher level individual. ERTAS has the following characteristics in common with higher level individuality:

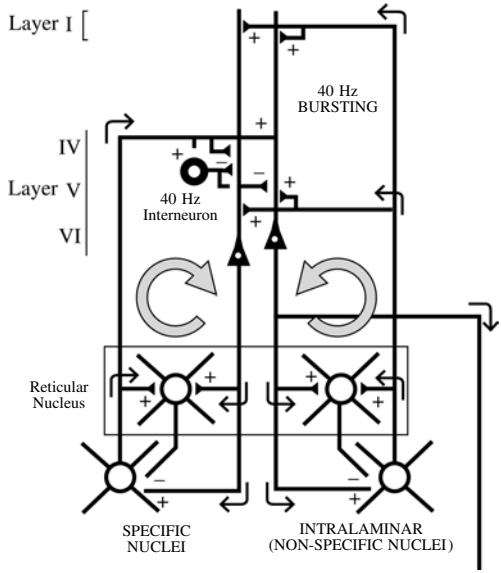


Figure 14.6 Thalamocortical circuits proposed to subserve temporal binding. Diagram of two thalamocortical systems. Left: Specific sensory or motor nuclei project to layer IV to the cortex, producing cortical oscillation by direct activation and feed-forward inhibition via 40-Hz inhibitory interneurons. Collaterals of these projections produce thalamic feedback inhibition via the reticular nucleus. The return pathway (circular arrow on the right) re-enters this oscillation to specific and reticularis thalamic nuclei via layer VI pyrimidal(*sic*) cells. Right: Second loop shows nonspecific intralaminary nuclei projecting to the most superficial layer of the cortex and giving collaterals to the reticular nucleus. Layer V pyramidal cells return oscillations to the reticular and the nonspecific thalamic nuclei, establishing a second resonating loop. The conjunction of the specific and nonspecific loops is proposed to generate temporal binding. (From Llinas et al. 1994, p. 260; as taken from Newman 1997)

1. It consists of *layers* of coordinated activity, each bound into a higher level of coordination, from the sweeping wave covering the cortex every 12-13 msec to the synchronously coherent 40-70 MHz oscillations between individual cortical layers of individual sensory systems within the wave to the intralayer communications between neurons in the RN and other places to the single-cell regulatory mechanisms of individual neurons.
2. High bandwidth positive and negative feedback channels manage the global state of the system at every level.
3. The global cortical state that emerges is the result of maximally compatible waves of activity, gated by the thalamus and RN.

4. The global state is an explicit selection of one of many possible states, based on some property of global coherence between wave contents.
5. Before cortical regions exhibit the coherent oscillations necessary for placing content into conscious states, the feedback mechanisms associated with the thalamus seem to gate, transform, and then broadcast input to neuronal layers throughout the cortex. By doing this, it seems to support a level of global constraint satisfaction consistent with the existence of a common receptivity for that activity.

As with the structure of vector coding that underlies the structure of phenomenal properties, the ERTAS outline of nature's solution to the production of conscious states is reassuringly coherent with the outline of what a higher level natural individual might look like "from the outside."

14.5 Summary

Part I of this book introduced some paradoxes, puzzles, and tensions that an adequate Liberal Naturalist theory of consciousness would have to solve, and the last chapter explained how the Consciousness Hypothesis handled those problems. This chapter went beyond those general conditions to look at how the entire framework, which is the Theory of Natural Individuals, might apply to our general understanding of related issues in philosophy, physics, and cognitive neuroscience. In philosophy it implies a strongly emergent nature for consciousness and explains how this nature is compatible with the claim that consciousness is neither epiphenomenal nor interactive. It also leads to a narrow kind of nonreductive functionalism. This chapter also explored some potential physical indications of natural individuality and further expanded the treatment of time put forward in chapter 10. The theory implies a kind of pluralism about time, dividing into S-time, which is fundamental and flows, and I-time, which is a logical construction from the relations of causal significance between causal processes that have their own distinct S-times. Finally, it examined two issues in cognitive neuroscience, pointing to ways our emerging understanding of the physical aspects of conscious states cohere with what the framework would lead us to expect if the consciousness hypothesis were true.

15

Conclusion

15.1 The Sliding Tile Puzzle Revisited

Consciousness is not a refugee. It simply has many homes. Like causation, it is at once of physics and metaphysics. Like experience, it is at once of philosophy and the cognitive sciences. It is a nexus for intellectual disciplines because it is a nexus within the world.

Chapter 1 ended with a promise to treat consciousness and causation like the last two pieces in a sliding tile puzzle. To keep that promise, I needed to challenge the established order in our picture of nature in a series of regressive moves and then reinstate the old order within a new context that was a more complete ordering of the entire puzzle. With the tiles properly ordered, I promised that consciousness and causation would have slid into their proper places and the established order in science will have returned.

I have made good on the metaphor of the sliding tile puzzle. By supporting arguments against physicalism, I challenged and rejected the chief scientific and philosophical hypothesis of the twentieth century, the hypothesis that the basic physical facts are all the basic facts. Having kicked away this metaphysical support, I faced up to a series of immediate problems that threaten to jumble our modern conception of our place in nature as natural objects. Two chapters argued at length that some kind of panexperientialism could be true and likely was true. Another chapter pointed out that experiencings could have had different boundaries than they do have, and it wondered how nature establishes the particular boundaries had by consciousness. Elsewhere, the book raised the issue of how to make sense of the idea that a single system could be both unified in the way consciousness seems to be and also fail to have that unity in the way that the brain seems to fail to have it. I worried about how to make sense of the subjective instant and how to avoid the dualist dilemma of having to choose between epiphenomenalism and interactionism. I paused to note the structural problems created

Figure 15.1 When the pieces of nature's sliding tile puzzle are put into order, it turns out that nature is more like a lush woodland vista than an austere desert landscape.

for our image of nature by the grain problem. Finally, I challenged the almost universal assumption that a predictively adequate physical theory is also causally complete and argued that physics does not provide a theory of causation.

Having made the regressive moves, I took a fresh look at the motley pieces of the jumbled sliding tile puzzle, proposing a strategy for returning its pieces to their old places of comfortable honor. My chosen strategy was to rethink causation itself. The complete framework turned out to be a *Theory of Natural Individuals*, and I began by focusing on what a causal constraint might be and proposing a theory of the causal nexus. Placing a modern twist on an ancient conception, I rebuilt the world as a mesh of interconnected individuals, each with effective and receptive aspects. Often using examples from an artificial physics, I modeled the metaphysical relations between effective and receptive properties in as much detail as possible. I named this first part the *Theory of Causal Significance*, and it was the foundation for the larger *Theory of Natural Individuals*.

The theory proposed that physical causation specifies only the regular and lawlike ways that the world's basic effective properties instantiate. This is the nomic mosaic. Because the existence of effective properties implies that individuals have receptive aspects or properties, to fully understand causation requires carefully modeling the receptive aspect of individuals. A novel element of the account was modeling receptivity as an irreducible connection that can build natural individuals at many levels of nature. Receptivity does the work of the causal connection.

These efforts led to the introduction of yet a third aspect of causation called *carriers*. Carriers are implementing elements for causal powers. I named this part of the theory of our world's causal nature *the Carrier Theory of Causation*. The carrier argument was that these effective and receptive sides of individuals are not pure dispositions but are *carried*. I gave arguments that nature's ultimate carriers must have structural and natural features that precisely mirror those had by experience. Experience and experiential qualities are kinds of carriers for the nomic content of natural individuals. If the models introduced are on the right track, the idea of understanding consciousness as a special kind of carrier, one involved in cognitive contexts, is very promising.

The duality of causal power is its division into effective constraints and the reception of these constraints. The duality of consciousness is its division into phenomenal quality and the experiencing of that quality. *The Consciousness Hypothesis* completed the *Theory of Natural Individuals*, by proposing that the duality of consciousness carries the duality of causation for a cognitively structured, high-level natural individual.

Within this new Liberal Naturalist framework, the intrinsic phenomenal qualities of consciousness are placed as carriers of effective constraints. The experiencing of these qualities is placed as the receptivity to these qualities belonging to a natural individual. Although consciousness cannot be understood on purely physical terms, it avoids both causal irrelevance and interactionism. The result is a panexperientialist view very reminiscent of the sort proposed by Whitehead (1929), in which the fundamental entities are processes composed of internally linked, experiential events. By introducing this analysis of causation, Liberal Naturalists find a place for consciousness. Have they restored the old order, though? Can science, and the scientific understanding of the world, survive within a Liberal Naturalism that endorses *the Theory of Natural Individuals*?

15.2 Science and the Carrier Theory

To wonder if science can survive the Theory of Natural Individuals is to wonder in the wrong direction. A more legitimate question is whether the natural individuals may survive an encounter with real science. The *Theory of Natural Individuals* is put forward as a general metaphysical analysis of causation. Any analysis of causation that makes the science of our world seem impossible or false must be rejected. I tried to develop the Theory of Natural Individuals as an articulation of some very general and intuitive principles explored within the framework of one toy physics or another. The real challenge for the Theory of Natural Individuals is this: Can it grow up? Can we identify the detailed physical conditions that indicate the receptive structure of the world in a way that provides operational criteria for deciding whether and where natural individuals exist? If we can, then we can test the theory by investigating whether the systems we know are conscious—the neural correlates of consciousness in ourselves and other mammals, for example—meet the criteria.

A theory that includes the facts about the world's receptive structure will contain physics as a proper subset, and that alone seems to guarantee that a finished carrier theory would preserve physical science in something essentially like its traditional place. The problem that really confronts us is to use our best science in an attempt to bootstrap our way into the total theory of the world's causal structure. This task might not be hopeless. Even though the world's receptive structure is strictly irrelevant to physical explanation, we can still make reasonable guesses about how to interpret physical theory in a way that involves receptivity. For instance, we have good reason to believe that the direction of time supervenes on asymmetric receptive connections. Although physical theory struggles to explain the direction of time, we nonetheless know in which direction it flows. We can reasonably postulate, given this knowledge, a receptive structure that is heavily asymmetric in that direction.

Additionally, certain puzzles in quantum mechanics, puzzles such as the EPR paradox, seem to betray the existence of receptive connections belonging to higher level individuals. We can examine the mathematical characteristics of such systems and try to extract the formal characteristics that betray such global properties (e.g., the existence of global constraints on the joint states of a system's components). Quantum coherent systems also are good candidates for being higher level individuals. Finally, the way that individuals *decohere* from one another will likely yield strong clues regarding the world's receptive structure. In examining this decoherence, it should prove fruitful to concentrate on multilevel interactions so as to get a picture of the stratified layers of our world's ontology and also to concentrate on systems in which we see causal amplification of effects from one layer to another.

All these suggestions are speculative, and they pose large technical problems. However, they constitute a possible research program, which is something more than Liberal Naturalists have had to this point. Should the program prove fruitful, the payoff would be worth the investment of effort. It would be nothing less than a simultaneous insight into some of the deepest truths about ourselves, about the world in which we are enmeshed, and about the metaphysical background of possibility from which we come.

If we have done today's work well, we may start looking forward to tomorrow's work. The challenges that lie ahead are exciting: A new world stands before us with a widened frontier. I would say that finding a place for consciousness is the easiest part of solving the hard problem. The ultimate prize is to understand the place we have come to in livable and workable detail.

Notes

Chapter 1

1. Baruss (1990) catalogs twenty-nine separate definitions of the term, which he groups into three categories. Chalmers (1996) distinguishes eight senses of the term in his first chapter.

2. These qualities provide *phenomenal information* to subjects of experience. For an extended defense of the existence of phenomenal information, see Lycan (1996). Lycan is a physicalist.

3. I do not mean to suggest in any way at all that these are independent capacities.

4. This is just an initial assumption, of course. For just about anything that one can name, it is not hard to find at least one philosopher who is willing to argue against its existence.

Chapter 2

1. "A priori" is a philosophical term for a conclusion that can be justified independently of an appeal to historical or scientific facts. "A posteriori" is the complementary term for conclusions that can be justified only by appealing to such facts.

2. A universal Turing machine is a kind of computer that can simulate any other computer.

3. Note that this entailment obviously holds even though gliders have not been defined in terms of *Life*'s physics.

4. More exact analyses of "conceptual," "empirical," and "interpretive" are given in chapter 3.

5. The relevant premise for the argument is (P): If x has the status of being an observable, then the evidence for x must also have the status of being observable. For example, if I can observe that it is cold outside based on the evidence that there is snow on the ground, the snow on the ground must also be something that I can observe. Similarly, if I observe a meson in the cloud chamber based on the evidence that a cloud track has appeared, then the cloud track must be an observable also. The premise (P) gains its plausibility from the principle that the epistemic status of evidence cannot be less secure than the status of that which it is evidence for.

Chapter 3

1. This chapter is a necessary evil relative to the main project in the book. It contains a lengthy and sometimes technical discussion of the disagreements surrounding the previ-

ous chapter's conclusion that the physical facts do not entail the facts about experience. Readers interested in those disagreements may benefit from reading this material, but those who are not philosophers or who feel comfortable with the argument in the previous chapter may wish to skip this chapter.

2. For example, a pure *Life* world without consciousness might seem possible but maybe not be "really" or "metaphysically" possible.

3. On some views of laws, they are just summations of already existing regularities in nature and so will not have any ontology of their own. I skip over that view here because I find it very problematic, and I argue against it at length in chapter 8, the first chapter of the section titled *The Faces of Causation*.

4. Reminder: I have begun using *entail* to mean a priori entail.

5. I owe thanks to David Chalmers for suggesting this as the best way to express my point here and in several other places in this section of the chapter.

6. Here I am treating information as something an ideal knower could discover.

7. I fully defend this requirement in section 3.3.4.

8. Appropriately enough, the section in Chalmers (1996) is titled "Almost everything logically supervenes on the physical" (p. 71).

9. Kirk's "Strict Implication Thesis" is equivalent to the logical supervenience requirement.

10. None of the cases helps the physicalist answer the antiphysicalist argument. The lack of entailment in the argument is not due to vagueness, so it would be no help to sharpen a boundary. The analysis gives a sharp account of "physical" and seems to preclude recognizing a previously unrecognized but sharp condition. Finally, we could simply move the boundary of our category "physical" to include facts that can entail facts about consciousness. This is akin to what the o-physicalist is proposing. However, that kind of decision to move a categorical boundary has to meet rational and social constraints on prior use of the term. I believe that, once the details of the proposed move are known, it is clear that it violates both the spirit and the letter of past use, and I have urged the o-physicalist to withhold judgment until understanding the details required to make the o-physicalist move work.

11. From this point forward, I am using "fact" to mean "narrow fact" unless I specifically say otherwise, similar to how I have been using "entailment" to mean "a priori entailment."

12. Perhaps along with some indexical specification.

13. For those suspicious of the a priori, section 3.4 discusses how to translate the arguments into a holist framework.

14. This framework also fits the arguments in chapter 2 by providing a basis for discussing them that bypasses some issues involved in the debate about conceivability arguments. For example, by defining entailment as an informational containment relation, my treatment bypasses many of the issues about the reference-fixing intensions on concepts that have driven some of the critiques of conceivability arguments such as the Zombie argument in Chalmers (1996). One such critique is in Hawthorne (2002), where Hawthorne produces a possible analysis of the reference-fixing conditions on our concept of consciousness. Hawthorne argues that his analysis is compatible both with Zombie intuitions and the truth of physicalism. He suggests that "consciousness" might be a kind of disjunctive concept: One disjunct gives reference fixing conditions if the world is wholly physical, and the other gives different reference fixing conditions if the world is not wholly physical. Hawthorne argues that the physicalist can respond to conceivability arguments by trading on uncertainty about which disjunct is reference-fixing given the totality of actually referred-to facts.

Unlike the Zombie argument that Hawthorne's critique aims at, the argument in chapter 2 is not a conceivability argument. A conceivability argument argues for a lack of entailment by arguing for the consistent conceivability of a world. The argument in chapter 2 inverts this: It argues for the consistent conceivability of a world (the *Life* world without consciousness) by directly arguing for a lack of entailment (phenomenal qualities cannot be analyzed into bare difference). The issue at stake there is accounting for certain observational information we have about consciousness. The information-based argument demands an account of the relations between the physical facts and all positive facts we possess about the phenomenal, making no claims about which and whether these facts go into determining reference. Therefore, even if the reference-fixing conditions on consciousness left the truth of physicalism open for us (as Hawthorne claims), physicalism would still fail if a purely physical world failed to ontologically necessitate these further positive facts we possess about consciousness. To see that the argument in chapter 2 stands irrespective of Hawthorne's analysis, we can simply note that the definition of a *pure Life* world included a totality fact of the type he suggests. Given that the totality fact did not enable facts about a pure *Life* world to entail facts about consciousness, the argument that facts about a pure physical world could not entail the facts about consciousness proceeds just the same.

15. Kripke writes, "Isn't the situation I just described also counterfactual? At least it may well be, if such Martians never in fact invade. Strictly speaking, the distinction I wish to draw compares how we *would* speak *in* a (possibly counterfactual) situation, *if* it obtained, and how we *do* speak *of* a counterfactual situation, knowing that it does not obtain—i.e., the distinction between the language we would have used in a situation and the language we *do* use to describe it." Here, Kripke arguably seems to be anticipating the Chalmersian point that the kind of counterfactual truths rigid designators produce are just one way of regarding possible situations that are otherwise accessible a priori, and so just one kind of modal truth, and do not represent ways of discovering substantial constraints on the set of possible worlds.

16. Some philosophers believe that the moral of Kripke's and Putnam's work is that facts about *identity*, not necessity, are what elude a priori entailment. The necessities come along as a trivial consequence of the identities. In the next subsection, I discuss the proposal that identity is the basis of a posteriori necessity. Here, it is enough to be clear that Kripke and Putnam did not discover a brand of essentialism that can do the physicalist's work. Whether or not there is an appropriate kind of identity is a different issue.

17. Would superempirical virtues such as simplicity give some reason to prefer lower-level determination? Certainly, but the reasons would be mixed and nonconclusive. Virtues such as simplicity have to be balanced against other virtues, such as expected fruitfulness. The history of discovery, such as Maxwell's discovery of the magnetic field equations, have taught us that reacting to a lack of entailment by searching for new fundamental laws and facts can be extraordinarily fruitful. If entailment is missing, there's really no nonsociological reason to prefer the potential simplicity of a lower-level determination story to the potential fruitfulness of a search for missing knowledge.

18. The controversial "identity of indiscernibles" states that indiscernible things are identical. It should not be confused with the uncontroversial principle of the "indiscernibility of identicals," which states that identical things are indiscernible.

19. Similarly, if there is a need to transfer properties from theory to commonsense entity, we conclude that our observation base is incomplete and consider designing experiments to test the consequence of the theory.

20. Although the consequent cannot contain *more* empirical information, in the sense that was defined, than the antecedent. See the beginning of this chapter.

21. When I say the primitive identity "results from" an a posteriori necessity, I mean that its primitiveness comes from the primitiveness of the supervenience relation underlying it.

22. As discussed in section 3.2, supervenience is a philosophical name for determination relations. Facts about *X* supervene on facts about *Y* if, and only if, the facts about *Y* determine the facts about *X*. For more discussion of the relation of supervenience to physicalism, see section 3.2. There is a large literature on different types of supervenience and their significance. Kim (1993) is a good starting point.

23. Indiscernibility proper also includes *modal indiscernibility*, which is natural indiscernibility in all possible worlds. Modal indiscernibility is relevant to classic identity puzzle cases such as Statue-vs-Lumpl, in which two things may be naturally indiscernible but not identical because they are not modally indiscernible. I ignore modal indiscernibility here because the case of consciousness hinges directly on natural indiscernibility.

24. Not including the property of being identical to each other.

25. By using an analogy to deflationism about truth to explain deflationism about identity, I do not mean to endorse deflationism about truth.

26. The argument is put in terms of token identity, but it can be run just as well against type identities.

27. The word *determine* is being used in an ontological sense. The exact kind of determination relation is left deliberately undefined.

28. These local high-level properties will not be identical to any of the properties explicitly involved in the *A*-facts, although they may be identical (and not just supervening) with facts determined by the *A*-facts.

29. Note that simplicity and other superempirical explanatory virtues can play no role here because they are not determination relations.

30. It could not be the other way around: that the *A*-facts entailed a higher-level property not entailed by the *B*-facts. *B* is the macrolevel point of view on the entity, and the *B*-facts contain all its macro properties just as the *A*-facts contain all the micro properties. If the *A*-facts entailed a fact that there was some higher-level property *P* where *P* was not in the set of *B*-facts, then there would be a straight contradiction between the lower- and higher-level facts, and the identity could not be true.

31. Because the point of view that delivers our phenomenal information is a system internal to *A* (by hypothesis), it is plausible to characterize the aspect of *P* revealed by it as a local aspect. We can then supplement our initial premises with an aspect-determination rule for use in the iterated argument: The *A*-facts determine all the local aspects of *A*.

32. To get the entailment, one has to bridge the gap between chemical theory and concepts such as *room temperature*. This may require putting some facts about empirical identities into the theoretical context, although usually mere realization will do. I dealt with these issues surrounding identity and necessity in the last section, and they do not provide a loophole in the account I give here.

33. Similar criteria are defended and formalized in Dunn (1973).

34. Although holism itself comes in varieties other than Quine's, I focus responding to his version and let my response to it serve as a template for responding to variants.

35. Searle describes "The Background" as a necessary grounding for intentionality that includes (at least) a network of nonrepresentational capacities, practices, and preintentional assumptions.

Chapter 4

1. I am using "quanta" with its classical meaning of "discrete units." I do not mean to suggest ties to quantum physics.

2. I think it presents the same kind of challenge even for nonreductive physicalists who were not persuaded by the arguments in chapters 2 and 3. It is an explanatory problem, and the explanatory problem arises just as strongly even if the possibilities used to generate it are just epistemic possibilities.

Chapter 5

1. This analogy problem can rightly be regarded as a promissory note. The note is paid in full in chapter 12, "The Carrier Theory of Causation," in which the kinds of similarities and differences at issue are described in more detail.
2. In chapter 3.

Chapter 6

1. Defined at the beginning of the last chapter.
2. People rarely endorse the position so baldly in print, but it is very common to hear it appealed to in an offhand way in conversation and informal correspondence.
3. There are other attempts to tie consciousness and intentionality nonreductively, as discussed briefly in the last chapter. These attempts are not biologically based, though, so they are not relevant to the discussion in this section of the chapter.

Chapter 7

1. Some readers will have an urge to reject altogether the problem of the subjective instant, but I ask for restraint at this point in the discussion.
2. The brain does seem to do bind contents (see chapter 14), but our reasons for believing it should bind contents predated the empirical discoveries and seem to rest on more primitive intuitions about our conscious states.
3. I am not here denying that phenomenal qualities have an intrinsic character. It is just that the existence of each quality within an experience seems to be tied together with the existence of the whole context despite its intrinsic character.
4. A "sufficient cause" is a cause that, by itself, can produce an event. A "probabilistic cause" is a cause that raises the probability than an event will happen, given some background conditions.
5. I am making the usual assumption that the randomness in QM is either a discovery that causation is probabilistic or a stopgap to be filled by hidden physical variables. I am purposely not treating it as a discovery that there are causal gaps in the physical world.
6. In his more recent book, Lycan (1996) takes a more deflationary attitude toward the problem.

Chapter 8

1. If current physics cannot meet this goal, then the article of faith is that an ideal future physics could. The faith implies that the measurement problem, for instance, can be resolved without appealing to nonphysical entities.
2. A "descriptively adequate characterization" names all the fundamental properties that feature in physical regularities, the laws governing their temporal evolution, and the laws governing their interactions.
3. Chapter 3 discusses the different aspects of this claim in detail.

4. Certain interpretations of quantum mechanics argue against this belief.

5. *Empiricism* is the view that knowledge of the external world comes from sense experience and sense experience alone. Chapter 3 discusses a variety of empiricism.

6. This formula should be read, "For all *x,* if *x* is an *F*, then *x* is a *G*."

7. This is true because of the way classical formal logic treats the logic of "*if . . . then . . .*" statements. They are deemed true whenever the antecedent, which falls between *if* and *then*, is false, regardless of whether the rest of the statement is true or false.

Chapter 9

1. Which notion, causal responsibility or causal significance, deserves the name *causation?* I think the ordinary language use of *causation* names causal responsibility, but I coopt the term for the rest of this book. In most places, when I use the term *causation*, I am talking about causal significance. In the few places where I use *causation* to mean causal responsibility, I hope that the context makes the switch clear. With luck, no harmful confusion will result from these slight equivocations.

2. And effective properties may perhaps even present constraints for the states of individuals previous to them, if the 4-D view of spacetime is correct.

3. Although I introduce a more technical and constrained notion later, here I follow Strawson (1959) in taking a liberal attitude toward the meaning of *individual*. An individual is simply an entity that bears properties. The reason for being so liberal is to avoid heavy commitments at this early stage to what the ultimate causal ontology will be like. To quote Strawson in full:

> So anything whatever can appear as a logical subject, an individual. If we define "being an individual" as "being able to appear as an individual," then anything whatever is an individual. So we have an endless variety of categories of individual other than particulars—categories indicated by such words as "quality," "property," "characteristic," "relation," "class," "kind," "sort," "species," "number," "proposition," "fact," "type," and so forth. (p. 227)

4. By this point scholars of causation will have realized that the theory of causal significance is going to be a theory of causal powers rather than a theory of natural law in the Dretske-Tooley-Armstrong (DTA) mode. As I continue to develop my view of the problem of causation, it should become clear that the central *problem* of causation, as I see it, is understanding the metaphysics of causal interaction: What purpose does causal interaction serve, and what are the grounds of Being that allow it to occur? I pass over the DTA model of natural laws because I believe it is not a very good model for gaining a deep understanding of causal interaction and so is not very useful for understanding the problem of causation developed in the text.

5. One might say that the determination problem is to the theory of causal significance what the problem of causal production is to the theory of causal responsibility.

6. For effective properties there are the interesting questions of whether and how new effective properties can emerge from the binding of existing effective properties. Because of the possibility of emergence, straight inheritance by the higher level individuals of bound effective properties from the lower level need not always occur. These issues turn out to be very interesting, and the principles of emergence for effective properties are discussed in detail in the next chapter.

7. And vice versa if R was also part of the completion of E_1.

8. If the effective entity in the binding relation is already complete then it need not take

up the nature of the incomplete receptivity, though the receptivity will still take up the effective property into its own nature as part of its own completion.

9. These two principles correspond roughly to the idea of concrescence in process philosophy. However, "seek" is not meant to have psychological connotations.

10. Here is my reasoning: The stability of particles implies that their constituent properties can hang together in a single nexus under a wide variety of circumstances. From that, we can infer that they do not constrain one another very much because, no matter what determinate values circumstances force the constituent properties to take on (within a wide range), they are able to remain together in the particle nexus. Also, the value of any property typically seems to be a function of the circumstances of the particle and not of the values of their fellow properties in the particle nexus, which implies that their fellow properties are not constraining them, at least not perceptibly.

11. Thanks to Anand Rangaranjan for suggesting that entropy might play a role here.

Chapter 10

1. It is a process in the same sense as the billiard balls from the last chapter. Processes like it are analogous but not identical to Whiteheadian processes.

2. Thanks to John Gregg for suggesting the summary in this paragraph.

3. Notice again the difference between causal laws and natural laws. In the example, one might be ignorant of the causal law, but still have possession of a natural law describing the oscillation as a lawful regularity.

4. Those familiar with the EPR experiments in quantum mechanics should notice the similarity between those experiments and the situation depicted here.

5. Lower level individuals linked together as a process, as $I_{1.2}$ and $I_{1.3}$ in diagram (g) are, present a minor complication. The shared constituent of the overlapping receptive connections acts as a mediating responsive link between the individuals in the process. Through this link, the set of possibilities for their joint instantiation is already constrained to a proper subset of their Cartesian product. Instead of the eight prior possibilities that would be available for the joint instantiation of two independent individuals such as $I_{1.2}$ and $I_{1.3}$, only four prior possibilities actually exist to present to $I_{2.1}$. The shrinkage in the prior possibility space occurs because, for the purposes of determining a possibility space, the two linked individuals need to be treated "as if" they formed a single three-member individual with a unique kind of constraint. Of course, they only form an "as if" individual, as the two individuals could, in fact, be carved off from one another by binding one but not the other within a higher level individual.

6. Epiphenomenal individuals are representable within the formalism, and I will produce a representation of one later, but I do not think they are allowable by the theory. The problem with epiphenomenal individuals is that the fundamental principles of the theory discussed in the last chapter, such as the principles that determination is completion and that individuals seek completeness, require that higher level individuals take lower level individuals as members only if there is some indeterminacy in their joint states and if the higher level individuals make those joint states more determinate. Those requirements would preclude the actual existence of an epiphenomenal higher-level individual.

7. This step of the argument does not apply in the limit case of pure deterministic causation. In the limit case, the causal power brought *e* from *C*. It could not have brought anything else from *C*, and so there is no need to appeal to a space of possibility, but that would not mean that it did not bring *e* from *C*.

8. According to Smolin in *Three Roads to Quantum Gravity*, the most promising approaches to the next generation of physics also take this approach.

Chapter 11

1. "Slot" is a metaphor for talking about the carrying capacity of a receptive connection. A receptive connection's number of slots is its carrying capacity, where its carrying capacity determines the number of individuals it can bind to.

2. "Ontological supervenience" was defined in chapter 3 as an umbrella term for any determination relation that provided an ontological free lunch between the base facts and the supervenient facts. The discussion in that chapter went on to argue at some length that a priori entailment was the only relation that could ground ontological supervenience.

3. Although it seems this may change when we have a theory of quantum gravity.

4. On an infinite grid, every cell would have a surrounding neighborhood with individuals like the ones depicted here, and there would be lots of overlap between such individuals and neighborhoods.

Chapter 12

1. When I speak of "categorical natures," I mean the kind of thing conveyed by an appropriate answer to the question, "What is it to be *X*?" for the property of being *X*, or "What is it to be an *X*?" for the property of being an *X*.

2. When making this claim, I do not wish to deny the importance of indexicality (i.e., designation) in fixing reference. Likely, physical concepts contain indexical components, as "electron" may express a rigid designator. As Daniel Stoljar has pointed out, electrons are arguably just the categorical natures that play the electron role in our world. The more important point is that, even if some categorical nature is picked out indexically by these concepts, the indexical place functions much like a variable in the conceptual structure. Even if the value of the index anchors the language system to categorical natures, and even if it does so in a way that depends on the deictic orientation of the concept user within its physical context, it is still functional roles that do the most essential work in fixing the *physical category* applied to that nature. The indexically designated nature is a nature that is otherwise *extrinsic* to these entities, relative to the system of physical concepts we employ in science. That is, these natures, if they exist, are extrinsic within the system of physical dispositions they are carrying.

3. This constitutes yet a third argument against physicalism, distinct from the failures of consciousness and causation to ontologically supervene. The facts about the natures of the carriers do not ontologically supervene, either.

4. Related arguments from the circularity/schematic nature of the physical to this conclusion are in Fales (1990), pp. 219–220, and Chalmers (1996), pp. 303–304.

5. Kneale (1949) also uses red/green incompatibility as an example of a *de re* incompatibility between properties, proposing that it might serve as an analogical model for relations between physical properties. My suggestion in the text goes beyond Kneale in several ways, chiefly in taking the panexperientialist step of suggesting that they may serve as more than merely a model, and also by proposing the more subtle variety of relations I discuss in the text.

6. This discussion is making partial payment on the promissory note at the end of chapter 5.

7. Not everyone agrees. See Lockwood (1989) for a different view.

Chapter 13

1. If formalized, it would be an axiom instantiated from the axiom schema suggested by the Central Thesis.

Chapter 14

1. Recall that each natural individual is a causal nexus.
2. It can be helpful to think of receptive connections as *causal infrastructure*, where the actual causing is being done by the effective properties. In a metaphor, effective properties are cars, causings are car crashes, and receptive connections are the roads.
3. I have been told that "cause" is not a truly accurate translation of Aristotle's word, but there is no exact translation in English, and "cause" is in the same family.
4. Notwithstanding this point, the determinateness of the high-level individual may not always imply the determinateness of lower levels. In some circumstances, indeterminate states of lower level individuals may support determinate states for higher level individuals. See chapter 10, section 5, the subsection titled "Reflections."
5. Proponents of a posteriori physicalism usually put their position in terms of "metaphysical supervenience." As discussed in chapter 3, I believe that terminology lacks clarity, but the meaning I gave to "ontological supervenience" at least left it initially open that metaphysical supervenience might be a kind of ontological supervenience.
6. The reference would be a variable value if the value of the effective property is indeterminate considered independently of the nexus, and the effective state of the individual to whom the property belongs is not *fixed* relative to the nexus (an individual is fixed relative to the nexus if it plays the nonreceiving role in an asymmetric connection; see figures 10.8 and 10.9.
7. The reference would be a constant value if the value of the effective property is determinate considered independently of the nexus.
8. In correspondence, Stapp says that the reduction of brain states into a *mixture* of many different states through decoherence is the core starting point of the many-worlds interpretation of quantum mechanics, as well as his own collapse view, and it is not considered controversial among physicists.

This page intentionally left blank

References

Akins, Kathleen. (1993) "A Bat without Qualities." In *Consciousness*, edited by M. Davies and G. W. Humphreys, 258–73. Cambridge, MA: Blackwell.

Armstrong, David. (1982) "Metaphysics and Supervenience." *Critica* 42, no. 14:3–17.

Armstrong, David. (1983) *What Is a Law of Nature?* New York: Cambridge University Press.

Baars, Bernard. (1997) "Commentary on Newman" Electronic seminar, Association for the Scientific Study of Consciousness, http://www.phil.vt.edu/ASSC/esem1.html.

Baruss, Imants. (1990) *The Personal Nature of Notions of Consciousness.* New York: University Press of America.

Block, Ned. (1980) "Troubles with Functionalism." In *Readings in the Philosophy of Psychology,* vol. 1, edited by N. Block, 268–306. Cambridge, MA: Harvard University Press.

Block, Ned. (1995) "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18, no. 2:227–47.

Block, Ned, and Robert Stalnaker. (1998) "Conceptual Analysis, Dualism, and the Explanatory Gap." *The Philosophical Review* 108, no. 1:1–46.

Bogen, J. E. (1995) "On the Neurophysiology of Consciousness," parts 1 and 2. *Consciousness and Cognition* 4:52–62; 4:137–58.

Boghossian, Paul. (1996) "Analyticity Reconsidered." *Nous* 30, no. 3:360-91.

Braude, Stephen E. (1991) *First Person Plural: Multiple Personality and the Philosophy of Mind.* New York: Routledge.

Brink, David. (1991) *Moral Realism and the Foundation of Ethics*. New York: Cambridge University Press.

Carnap, R. (1928) *The Logical Construction of the World*. Berlin: Weltkreis. Translated by R.A. George. Berkeley and Los Angeles: University of California Press, 1967.

Carnap, R. (1937) *The Logical Syntax of Language.* Translated by A. Smeaton. London: K. Paul, Trench, Trubner.

Chalmers, David. (1994) "The Components of Content." Unpublished manuscript.

Chalmers, David. (1995) "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2, no. 3:200–219.

Chalmers, David. (1996) *The Conscious Mind*. New York: Oxford University Press.

Chalmers, David. (1997) "Moving Forward on the Problem of Consciousness." In *Explaining Consciousness: The Hard Problem,* edited by Jonathan Shear, 379–422. Cambridge, MA: MIT Press.

Chalmers, David. (2002) "Does Conceivability Entail Possibility?" In *Imagination, Conceivability, and Possibility*, edited by T. Gendler and J. Hawthorne. New York: Oxford University Press.

Chalmers, David. (2004) "The Foundations of Two-Dimensional Semantics." In *Two-Dimensional Semantics: Foundations and Applications*, edited by M. Garcia and J. Macsa. New York: Oxford University Press.

Chisholm, Roderick M. (1955) "Law Statements and Counterfactual Inference." *Analysis* 15, no. 5:97–105.

Chisholm, Roderick M. (1966) "Freedom and Action." In *Freedom and Determinism*, edited by Keith Lehrer. New York: Random House.

Churchland, Patricia S., and Paul M. Churchland. (1990) "Intertheoretic Reduction: A Neuroscientist's Field Guide." *Seminars in the Neurosciences,* 2, 249–256.

Collingwood, R. G. (1940) *An Essay on Metaphysics*. Oxford: Clarendon Press.

Crick, Francis. (1994) *The Astonishing Hypothesis*. New York: Simon and Schuster.

Crick, F., and Koch, C. (1990) Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences* 2:263–75.

Crick, F., & Koch, C. (1995) "Are We Aware of Neural Activity in Primary Visual Cortex?" *Nature* 375:121–23

Cytowic, R. E. (1989) *Synesthesia: A Union of the Senses*. New York: Springer-Verlag.

Cytowic, R. E. (1993) *The Man Who Tasted Shapes: A Bizarre Medical Mystery Offers Revolutionary Insights into Reasoning, Emotion, and Consciousness.* New York: Putnam.

Cytowic, R. E. (1995) "Synesthesia: Phenomenology and Neuropsychology: A Review of Current Knowledge." *Psyche* 2, no. 10: http://psyche.cs.monash.edu.au/v2/psyche 2-10-cytowic.html.

Das, Rajarshi, Melanie Mitchell, and James Crutchfield. (1994) "A Genetic Algorithm Discovers Particle Based Computation in a Cellular automata." In *Parallel Problem Solving in Nature,* vol. 3 of Lecture Notes in Computer Science, edited by Y. Davidov, H.-P, Schwefel and R. Manner, 344–53. Berlin: Springer-Verlag.

Davidson, Donald. (1967) "Causal Relations." *Journal of Philosophy* 64:426–41.

Dennett, Daniel C. (1988) "Quining Qualia." In *Consciousness in Contemporary Science,* edited by A. Marcel and E. Bisiach, 42–77. Oxford: Clarendon Press.

Dennett, Daniel C. (1991a) *Consciousness Explained*. Boston, MA: Little, Brown.

Dennett, Daniel C. (1991b) "Real Patterns." *Journal of Philosophy* 88, no. 1:27–51.

Dretske, Fred. (1986) "Misrepresentation." In *Belief: Form, Content, and Function,* edited by R. Bogdan. New York: Oxford University Press.

Dretske, Fred. (1995) *Naturalizing Mind*. Cambridge, MA: MIT Press.

Ducasse, C. J. (1926) "On the Nature and Observability of the Causal Relation." In *Causation,* edited by Ernest Sosa and Michael Tooley, 125–136. New York: Oxford University Press.

Dunn, Michael J. (1973) "A Truth Value Semantics for Modal Logic." In *Truth, Syntax, and Modality*, edited by H. Leblanc. Amsterdam: North-Holland.

Edelman, Gerald. (1985) "Neural Darwinism: Population Thinking and Higher Brain Function." In *How We Know*, edited byMichael Shafto. San Francisco, CA: Harper & Row.

Edelman, Gerald. (1989) *The Remembered Present.* New York: Basic Books.

Fales, Evan. (1990) *Causation and Universals.* New York: Routledge.

Flanagan, Owen. (1992) *Consciousness Reconsidered.* Cambridge, MA: MIT Press.

Gray, J. A. (1996) "The Contents of Consciousness: A Neuropsychological Conjecture." *Behavioral and Brain Sciences* 18, no. 4:659–722.

Griffin, David Ray. (1997) "Panexperientialist Physicalism and the Mind-Body Problem." *Journal of Consciousness Studies* 4, no. 3:248–68.

Griffin, David Ray. (1998) *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*. Berkeley: University of California Press.

Güzeldere, Güven. (1997) "The Many Faces of Consciousness: A Field Guide." In *The Nature of Consciousness: Philosophical Debates,* edited by Ned Block, Owen Flanagan, and Güven Güzeldere, 1–68. Cambridge, MA: MIT Press.

Hamilton, Edith, and Huntington Cairns, eds. (1961) "The Sophist." In *Plato: Collected Dialogues.* Princeton, NJ: Princeton University Press.

Harré, R., and E. H. Madden. (1975) *Causal Powers: A Theory of Natural Necessity*. Oxford: Blackwell.

Haugeland, John. (1993) "Pattern and Being." In *Dennett and His Critics*, edited by Bo Dahlbom, 53–69. Oxford: Blackwell.

Hawthorne, John. (2002) "Advice for Physicalists." *Philosophical Studies* 108:17–52.

Hill, Christopher S. (1997) "Imaginability, Conceivability, Possibility and the Mind-Body Problem." *Philosophical Studies* 87:61–85.

Hodgson, David. (1991) *The Mind Matters*. Oxford: Clarendon Press.

Horgan, Terence. (1984) "Supervenience and Microphysics." *Pacific Philosophical Quarterly* 63: 29–43.

Hume, David. (1748) *Inquiry Concerning Human Nature.* Edited by Anthony Flew. 1962: 17–164.

Jackson, Frank. (1982) "Epiphenomenal Qualia." *Philosophical Quarterly* 32, no. 127:127–36.

Jackson, Frank. (1986) "What Mary Didn't Know." *Journal of Philosophy* 83, no. 5:291–95.

Jackson, Frank. (1994) "Armchair Metaphysics." In *Philosophy in Mind*, edited by Michaelis Michael and John O'Leary-Hawthorne, 23–42. Dordrecht: Kluwer.

Jackson, Frank. (1997) "Finding Mind in the Natural World." In *The Nature of Consciousness: Philosophical Debates.* Edited by Ned Block, Owen Flanagan, and Güven Güzeldere, 483–91. Cambridge, MA: MIT Press.

Johnston, Mark. (1992) "Constitution Is Not Identity." *Mind* 101, no. 401:89–105.

Johnston, Mark. (1996) "It Necessarily Ain't So." Unpublished manuscript.

Kafatos, Menas, and Robert Nadeau. (1990) *The Conscious Universe: Part and Whole in Modern Physical Theory*. New York: Springer-Verlag.

Katz, Jerrold J. (1990) *The Metaphysics of Meaning*. Cambridge, MA: MIT Press.

Kim, Jaegwon. (1982) "Psychophysical Supervenience." *Philosophical Studies* 41, no. 1:51–70.

Kim, Jaegwon. (1984) "Epiphenomenal and Supervenient Causation." In *Midwest Studies in Philosophy IX: Causation and Causal Theories*, edited by Peter A. French et al., 257–270. Minneapolis: University of Minnesota Press.

Kim, Jaegwon. (1993) "Supervenience and Mind." Cambridge: Cambridge University Press.

Kim, Jaegwon. (2000) *Mind in a Physical World.* Cambridge, MA: The MIT Press.

Kirk, Robert. (1974) "Zombies versus Materialists." *Aristotelian Society* 48 (suppl.):135–52.

Kirk, Robert. (1994) *Raw Feeling.* Oxford: Clarendon Press.

Kirk, Robert. (1995) "How Is Consciousness Possible?" In *Consciousness and Experience,* edited by Thomas Metzinger, 391–407. Thorverton, UK: Imprint Academic.

Kneale, W. C. (1949) *Probability and Induction.* Oxford: Oxford University Press.

Koch, Christof. (1996) "Towards the Neuronal Substrate of Visual Consciousness." In

Towards a Science of Consciousness: The First Tucson Discussions and Debates, edited by S. R. Hameroff, A. W. Kaszniak, and A. C. Scott. Cambridge, MA: MIT Press.

Kripke, Saul. (1971) "Identity and Necessity." In *Meaning and Reference*, edited by A. W. Moore, 162–91. New York: Oxford University Press.

Kripke, Saul. (1972) "Naming and Necessity." In *Semantics of Natural Language,* edited by D. Davidson and G. Harman, 254–355. Dordrecht, Netherlands: Kluwer.

LaBerge, David. (1995) *Attentional Processing: The Brain's Art of Mindfulness.* Cambridge, MA: Harvard University Press.

Leopold, D. A., and N. K. Logothetis. (1996) "Activity Changes in Early Visual Cortex Reflect Monkeys' Percepts during Binocular Rivalry." *Nature* 379:549–53.

Levine, Joseph. (1993) "On Leaving Out What It's Like." In *Consciousness*, edited by M. Davies and G. W. Humphreys, 121–36. Cambridge, MA: Blackwell.

Levine, Joseph. (1995) "Qualia: Intrinsic, Relational, or What?" In *Conscious Experience*, edited by Thomas Metzinger, 277–92. Thorverton, UK: Imprint Academic.

Levine, Joseph. (1998) "Conceivability and the Metaphysics of Mind." *Nous* 32, no. 4:449–80.

Lewis, David. (1973) "Causation." *Journal of Philosophy* 70, no. 17:556–71.

Lewis, David. (1981) "Are We Free to Break the Laws?" *Theoria* 47, no. 3:113–21.

Lewis, David. (1983) "Extrinsic Properties," *Philosophical Studies* 44:197–200.

Lewis, David. (1988) "What Experience Teaches," *Proceedings of the Russellian Society*, University of Sydney, Australia. Reprinted in *Mind and Cognition: A Reader*, edited by William Lycan (Cambridge, MA: Blackwell, 1990), 490–498.

Lewis, David. (1995) "Should a Materialist Believe in Qualia?" *Australasian Journal of Philosophy* 73, no. 1:140–44.

Llinas, R., and D. Pare (1996) "The Brain as a Closed System Modulated by the Senses." In *The Brain-Mind Continuum*, edited by R. Llinas and P. S. Churchland. Cambridge, MA: MIT Press.

Llinas, R., et al. (1994) "Content and Context in Temporal Thalamocortical Binding." In *Temporal Coding in the Brain*, edited by G. Busaki et al. Heidelberg, Germany: Springer-Verlag.

Loar, Brian. (1990) "Phenomenal States." *Philosophical Perspectives* 4:81–108.

Locke, John. (1690) *An Essay Concerning Human Understanding.* Edited by Peter H. Nidditch, 1975. Oxford: Clarendon Press.

Lockwood, Michael. (1989) *Mind, Brain, and Quanta*. Cambridge, MA: Blackwell.

Lockwood, Michael. (1993) "The Grain Problem." In *Objections to Physicalism*, edited by Howard Robinson, 271–91. Oxford: Clarendon Press.

Logothetis, N., and J. Schall. (1989) "Neuronal Correlates of Subjective Visual Perception." *Science* 245:761–63.

Lycan, William G. (1987) *Consciousness*. Cambridge, MA: MIT Press.

Lycan, William. (1990) "The Continuity of Levels of Nature." In *Mind and Cognition: A Reader*, edited by William Lycan, 80–96. Cambridge, MA: Blackwell.

Lycan, William G. (1996) *Consciousness and Experience.* Cambridge, MA: MIT Press.

Marks, Charles E. (1981) "Commissurotomy, Consciousness, and Unity of Mind." Cambridge, MA: MIT Press.

Maxwell, Grover. (1971) *Structural Realism and the Meaning of Theoretical Terms*. Minnesota Studies in the Philosophy of Science, edited by M. Radner and S. Winokur, S. Vol. 4. Minneapolis: University of Minnesota Press.

Maxwell, Grover. (1979) *Rigid Designators and Mind-Brain Identity.* Minnesota Studies in Philosophy of Science, edited by C. W. Savage. vol. 9. Minneapolis: University of Minnesota Press.

McGinn, Colin. (1989) "Can We Solve the Mind-Body Problem?" *Mind* 98:349–66.

McGinn, Colin. (1993) "Consciousness and Cosmology: Hyperdualism Ventilated." In *Consciousness*, edited by M. Davies and G. W. Humphreys, 155–77. Cambridge, MA: Blackwell.

McGinn, Colin. (1995) "Consciousness and Space." In *Conscious Experience*, edited by Thomas Metzinger, 149–164. Imprint Academic.

Mellor, D. H. (1995) *The Facts of Causation.* New York: Routledge.

Metzinger, Thomas. (1995) "Faster than Thought: Holism, Homogeneity, and Temporal Coding." In *Conscious Experience*, edited by Thomas Metzinger, 425–64. Thorverton, UK: Imprint Academic.

Millikan, Ruth G. (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, Ruth G. (1993) *White Queen Psychology and Other Essays for Alice.* Cambridge, MA: MIT Press.

Nagel, Thomas. (1974) "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4:435–50.

Nagel, Thomas. (1986) *The View from Nowhere.* New York: Oxford University Press.

Nagel, Thomas. (1998) "Conceiving the Impossible and the Mind-Body Problem." *Philosophy* 73, no. 285:337–52.

Nemirow, Laurence. (1990) "Physicalism and the Cognitive Role of Acquaintance." In *Mind and Cognition: A Reader*, edited by William Lycan, 490–98. Cambridge, MA: Blackwell.

Newman, J. (1995) "Reticular-Thalamic Activation of the Cortex Generates Conscious Contents," *Behavioral and Brain Sciences* 18, no. 4:691–92.

Newman, James. (1997) "Thalamocortical Foundations of Conscious Experience," Electronic seminar, Association for the Scientific Study of Consciousness, http://www.phil.vt.edu/ASSC/esem1.html.

Newman, James, Bernie Baars, et al. (1997) "A Neural Global Workspace Model for Conscious Attention." *Neural Networks* 10, no. 7:1195–1206.

Papineau, David. (1993) "Physicalism, Consciousness, and the Antipathetic Fallacy." *Australasian Journal of Philosophy* 71:169–83.

Penrose, Roger. (1989) *The Emperor's New Mind*. New York: Oxford University Press.

Poland, Jeffrey. (1994) *Physicalism*. Oxford: Clarendon Press.

Poundstone, William. (1985) *The Recursive Universe.* New York: William Morrow.

Price, Huw (1996) *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. New York: Oxford University Press.

Putnam, Hilary. (1973) "Meaning and Reference." *Journal of Philosophy* 70, no. 19:699–711.

Quine, W. V. (1963) "Two Dogmas of Empiricism." In *From a Logical Point of View*. New York: Harper & Row.

Quine, W. V. (1992) *The Pursuit of Truth*. Cambridge, MA: Harvard University Press.

Rosenberg, Gregg, Anderson, Michael. (2004) "Content and Action: The Guidance Theory of Representation." Unpublished manuscript.

Russell, B. (1927) *The Analysis of Matter*. London: Kegan Paul.

Salmon, Wesley. (1980a) "Probabilistic Causality." *Pacific Philosophical Quarterly* 61, no. 1:50–74.

Salmon, Wesley. (1980b) "Causality: Production and Propagation." In *Causation,* edited by Ernest Sosa and Michael Tooley, 154–71. New York: Oxford University Press.

Salmon, Wesley (1984) *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

Sanford, David. (1984) "The Direction of Causation and the Direction of Time." In *Midwest Studies in Philosophy IX: Causation and Causal Theories*, edited by Peter A. French et al., 53–75. Minneapolis: University of Minnesota Press.

Scott, Alwyn. (1995) *Stairway to the Mind*. New York: Springer-Verlag.

Searle, John. (1983) *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Searle, John. (1992) *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sellars, Wilfrid. (1963a) "Philosophy and the Scientific Image of Man." In *Science, Perception, and Reality*. New York: Humanities Press.

Sellars, Wilfrid. (1963b) "Some Reflections on Language Games." In *Science, Perception, and Reality*. New York: Humanities Press.

Shannon, C.E. (1948) "A Mathematical Theory of Communication." *Bell Systems Technical Journal* 27:379-423. [Reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Urbana: University of Illinois Press, 1949]

Shoemaker, Sydney. (1975) "Functionalism and Qualia." *Philosophical Studies* 27, no. 5:291–315.

Shoemaker, Sydney. (1996) *Causal and Metaphysical Necessity*. Unpublished manuscript.

Sydney Shoemaker. (1999) "On David Chalmers' *The Conscious Mind*." *Philosophy and Phenomenological Research* 59:539-44.

Siewart, Charles Peter. (1998) *The Significance of Consciousness.* Princeton, NJ: Princeton University Press.

Smart, J. J. C. (1959) "Sensations and Brain Processes." *Philosophical Review* 68:141–56.

Smart, J. J. C. (1993) "Laws of Nature as Species of Regularity." In *Ontology, Causality, and Mind: Essays in Honour of David M. Armstrong,* edited by John Bacon et al. Cambridge: Cambridge University Press.

Smith, B. C. (1996) *On the Origin of Objects*. Cambridge, MA: MIT Press.

Smolin, Lee. (2001) *Three Roads to Quantum Gravity.* New York: Basic Books.

Sprigge, T. L. S. (1994) "Consciousness." *Synthese* 98:73–93.

Stapp, Henry. (1996a) "The Hard Problem: A Quantum Approach." *Journal Of Consciousness Studies* 3, no. 3:194–210.

Stapp, Henry. (1996b) "Chance, Choice, and Consciousness: A Causal Quantum Theory of the Mind/Brain." Unpublished manuscript.

Stapp, Henry. (2000) *The Importance of Quantum Decoherence for Brain Processes* (*Response to Tegmark's paper*), http://www-physics.lbl.gov/~stapp/stappfiles.html.

Stoljar, Daniel. (2001) "Two Conceptions of the Physical." *Philosophy and Phenomenological Research* 62, no. 2:253–81.

Strawson, Galen. (1999) "Realistic Monism." In *Chomsky and His Critics*, edited by L. Antony and N. Hornstein. Oxford, UK: Blackwell.

Strawson, P. F. (1959) *Individuals.* London: University Paperbacks.

Tooley, Michael. (1987) *Causation: A Realist Approach.* Oxford: Oxford University Press.

Tye, Michael. (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* Cambridge, MA: MIT Press.

Tye, Michael. (1996) "The Function of Consciousness." *Nous* 30, no. 3:287–305.

Tye, Michael. (2000) *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Van Gulick, Robert. (1988) "A Functionalist Plea for Self-Consciousness." *Philosophical Review*, no. 97:149-188.

Van Gulick, Robert. (1993) "Understanding the Phenomenal Mind: Are We All Just Armadillos?" In *Consciousness*, edited by M. Davies and G. W. Humphreys, 137–154. Cambridge, MA: Blackwell.

Vendler, Zeno. (1962) "Effects, Results, and Consequences," *Analytical Philosophy*, edited by Ronald J. Butler, 1–15. Oxford: Oxford University Press.

Wager, Adam. (1999) "The Extra Qualia Problem: Synaesthesia and Representationism." *Philosophical Psychology* 12, no. 3:263–81.

Weiskrantz, Lawrence. (1986) *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press.

Weiskrantz, Lawrence. (1988) "Some Contributions of Neuropsychology of Vision and Memory to the Problem of Consciousness." In *Consciousness in Contemporary Science,* edited by A. Marcel and E. Bisiach, 183–99. Oxford: Clarendon Press.

Whitehead, Alfred North. (1929) *Process and Reality* (corrected edition), edited by Donald Ray Griffin and Donald W. Sherburne, 1978. New York: Free Press.

Wilkes, Kathleen. (1988) "Yishi, Duh, Um, and Consciousness." *Consciousness in Contemporary Science,* edited by A. Marcel and E. Bisiach, 16–41. Oxford: Clarenden Press.

Wolfram, Stephen. (2002) *A New Kind of Science.* Champaign, IL: Wolfram Media.

Yablo, Stephen. (1987) "Identity, Essence, and Indiscernability." *Journal of Philosophy* 84, no. 6:293–314.

Yablo, Stephen. (1993) "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53, no. 1:1–42.

Yablo, Stephen. (1993) "Cause and Essence." *Synthese* 93, no. 3:403–449.

Yablo, Stephen. (1999) "Concepts and Consciousness: Comments on Chalmers, The Conscious Mind." *Philosophy and Phenomenological Research* 59:455–464.

This page intentionally left blank

Index

Numbers in bold indicate defined terms.